1-1-2009

Multicast Optimization And Recovery In Multihoming Environment

Frank Levstek
Ryerson University

MULTICAST OPTIMIZATION AND RECOVERY IN MULTIHOMING ENVIRONMENT

by

Frank Levstek
B. Eng. Computer Engineering, Ryerson University, Canada, 2006

A thesis
presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering

Toronto, Ontario, Canada, 2009

Author's Declaration

I hereby declare that I am the sole author of this thesis or dissertation.
I authorize Ryerson University to lend this thesis or dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis or dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Borrower's Page

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Name	Signature	Address	Date

ABSTRACT

MULTICAST OPTIMIZATION AND RECOVERY IN MULTIHOMING ENVIRONMENT

Master of Applied Science
Department of Electrical and Computer Engineering
Ryerson University

Reliability of multicasting is increasingly becoming an important issue as the number of end users continues to grow, their demand for reliable service increases. This thesis proposes a novel algorithm for creating a recovery model while optimizing both inter and intra domain bandwidth. This is achieved by creating a centralized rendezvous point within the intra domain topology. The rendezvous point will create a static multicast tree and it will avoid link congestion during inter-domain link failure. This algorithm also reduces link congestion surrounding the border routers. This is achieved by shifting the root of the multicast tree from the border router to the rendezvous point. This rendezvous point is then selected based on an optimization algorithm to reduce bandwidth congestion. A Steiner tree was used to optimize the intra domain links. The simulation results indicate up to 30 % increase over conventional optimization algorithms which do not consider a rendezvous point model.

Keywords: *multicast, multi-homed, fault anomaly detection, bandwidth optimization*

Acknowledgments

I would like to thank my supervisor, Dr. Jaseemuddin for all the support and help. He saw potential in me and provided great intellectual and moral support.

I would like to thank my parents for always looking out for my best interest. They supported me, and were patient during my struggles and mistakes.

Table of Contents

Table of Figures

Chapter 1: Introduction

Many of today's content multimedia distribution companies are changing their distribution system to a digital platform. This digital platform offers many advantages to the end user. These advantages include greater and user controlled selection of content, digital sound and enriched content, more configuration and replay options for devices of variety of form factors, and other various improvements. For the provider it can leverage network engineering to reduce operating costs and allow for better competition. Another advantage for the provider is that the content can be distributed and repackaged over a large network.

Traditional digital distribution methods through digital networks do not scale well in large scale applications. Distribution using unicast protocol of User Datagram Protocol (UDP) is only efficient in small scale applications. This is due to the large single bandwidth requirements of sending duplicate data packets to the same branch of end users. For large scale applications, the preferred protocol for distribution to the end users is by the use of the multicast routing. In multicast routing, multicast distribution tree is laid out from the source to all the connected receivers. The use of multicast removes the requirement of sending individual packets to each end user that requires the same content. Instead one packet is sent down through the network. When the router close to the branching point of the multicast distribution tree receives this packet it duplicates the packet and distribute it to all end receivers.

The general premise of multicast is simple, and as a distribution means it provides excellent scalability. These multicast streams will only contain one type of packet which can be defined by group address. By defining different group addresses multiple streams can be defined. This will allow for multiple content streams to be distributed simultaneously. One such example of this would be having multiple video streams of different channels. The advantage of managing video content in this manner is that the

content can be distributed to multiple providers which can be organized in hierarchy of provider networks.

The streaming application is currently taking shape in the Digital TV and broadcast industry and is growing in popularity. As the technology becomes more widely accepted there is an increased need and interest in the reliability factor. Customers and content providers demand near perfection in all aspects. Advertising during special events is critical and outages are not acceptable. These requirements are transforming multicast distribution and routing to become more reliable, which in a worst case scenario must able to recover almost instantly with no or very low packet losses thereby avoiding jitter in human perception.

Multicast can experience a multitude of different scenarios which can contribute to loss of packets. To the end user they will experience either video artifacts or loss of signal. The work discussed in this thesis is aimed at reducing disruption of service while optimizing inter and intra domain bandwidth. While this is just one application of multicast it is the dominant use of it. Packet size and video compression in combination with multicast distribution also play a pivotal role in the fault scenario. While packet size is not directly analyzed in this thesis there is a correlation of multicast traffic and its effect in a fault scenario. As compression increases any slight loss in the stream is magnified. All these factors contribute to a disruption.

This thesis defines an algorithm to minimize system stress within a network topology. The algorithm which has been defined as Single Rendezvous Point (SRP) will optimize links within a topology and reduce link congestion. SRP creates a static optimized tree below the Rendezvous Point (RP). The RP acts as a central point in distributing a multicast stream to the receivers within a topology. If a link failure occurs upstream from the RP, bandwidth conservation will still be maintained. This is attributed to the RP acting as a root router when distributing the multicast stream.

1.1 Problem Definition

1.1.1 Overview

In the use of multimedia applications, multicast routing can be efficiently used to distribute content throughout a network. One major concern with multicast routing is the recovery process, which can be slow depending on its interaction with functions within the network layer and the layers below. Two popular solutions for multicast are PIM-SM and PIM-SSM; these rely on the unicast routing table. When a failure occurs in the network such as if a link breaks or a router goes offline, the unicast routing undergoes reconvergence. This process in turn triggers reconvergence event in the multicast routing. The reconvergence trigger in the multicast routing causes the multicast to start the Reverse Forward Path (RFP) algorithm which determines the path back to the source. The problem incurred with this procedure is that RFP check is based on input of the received multicast packet. If this interface is down, providing that the unicast routing table has not been updated, it may cause a delay in the reconvergence. This delay is compounded because of the coupling between the unicast and the multicast routing. The multicast routing protocol can only be able to reconverge once the unicast routing table has been reestablished.

1.1.2 Impact & Structure

As the multicast reconvergence occurs the user experiences either artifacts or a lost signal. Even small outages that may occur for a few hundred milliseconds may appear as a lost signal. Lab results have shown that outages as short as 20 to 30 milliseconds appear as artifacts. In Table1.1 a time breakdown of different outages and their outcomes can be seen [1].

Light Visual impacts [s]	Noticeable visual impact [s]	"Channel unavailable" message [s]
0.03 to 1	1 to 3	3.5 +

Table 1.1: Effects of time outages [1]

Assessing the impact is quite difficult because it depends where the outage occurs. If the outage occurs closer to the content provider the results would be catastrophic. The possible number of users affected by such an outage could be all currently subscribed users. On the contrary if the link is severed closer downstream towards the end user only a small number of users would be affected.

There are mechanisms in place to provide recovery. But these mechanisms are not adequate for seamless transmission. The other problem that arises is that network administrators have implied certain QoS measures to control the flow of traffic. These measures provide beneficial attributes in directing traffic but can cause added delay to the recovery time based on setup [2].

1.2 Thesis layout

Chapter two details the vital background information. This background information will provide an understanding of bandwidth optimization and how it is related to network anomalies in a multihomed topology. When optimization is implemented it can provide a means of controlling link failures and bandwidth constraints. By optimizing a path for a multicast stream bandwidth improved conversation can be achieved. Modeling the topology as a Steiner tree allows for reduced links between the source and the receivers. This, is in comparison of using a shortest path tree, which does not optimize link usage. The Steiner tree can be overlaid on the existing unicast routing table.

Chapter three discusses two different models the newly proposed Single Rendezvous Point (SRP) and Border router Rendezvous Points (BRP). SRP is an optimization algorithm used to create a centralized RP between two different border routers. SRP will

create an optimized path from the centralized RP to the receivers using a Steiner tree model. BRP performs the same optimization but assumes that each border router is a RP. Chapter three outlines the key differences between these two different optimizations.
Chapter four details the performance gains between SRP and BRP. SRP does have performance advantages over BRP but are based on the environmental conditions. SRP performs better in denser receiver environments which are supported by the results. Chapter four also postulates the performance gains and how they attributed to network performance and bandwidth conservation.

These chapters will build a comprehensive understanding of how SRP can be used to deal with routing anomalies and bandwidth conservation. They outline in detail how SRP can be applied to a network topology and discuss its efficiency. SRP does have certain limitations and positive aspects. All these attributes have been analyses and critiqued which will provide an understanding of the SRP optimization.

Chapter 2: Background

This section will discuss typical topology setups and other distribution methods throughout different network topologies. It will also give insight in how particular tools can be used to identify routing anomalies in the unicast routing table. These methods are important in identifying problems and can be used in initiating a fast recovery algorithm.

2.1 Distribution topology

Distribution of IPTV is a growing trend which more content providers are investing in. Deployment costs are decreasing due to the expansion of high speed internet services. The end user only requires a high speed internet connection and a set top box (STB). This STB will typically be connected to the end users home network which will use a router as a gateway to the provider's network.

This high speed internet connection for the end user can either be DSL, Cable, or even Fiber to the home. This connection link is described in figure 2.1 below. The access node in the figure would be the local loop of aggregation devices. This would be the last hop before a direct run to the home user. From this point the access node would receive content from the aggregation network. This aggregation network would contain a multitude of different access nodes. The aggregation network listed in figure 2.1 could represent a local area such as a city or smaller region depending on density.

The distribution network could be described as a local service provider. The local service provider would use their network to distributed video streams to their end users. The last part of the distribution tree is the video head end where video content is injected into the network [1]. As video information traverses throughout the network from the Video Head End it disperses more widely. The advantage of such a topology is that channels which can be represented as multicast streams can be redistributed through other distribution

networks. This allows other service providers to re-brand channels and sell them to the end user.

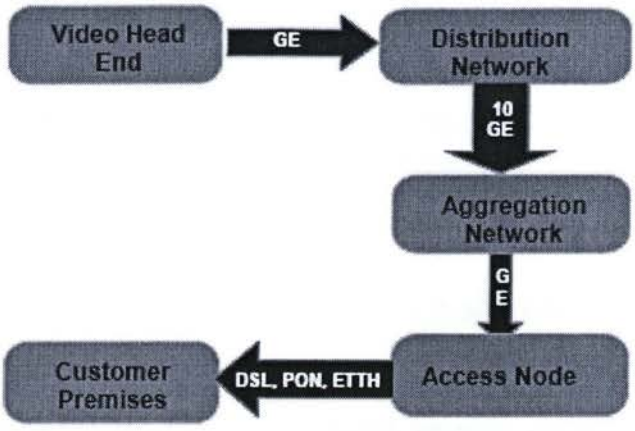

Figure 2.1: Distribution Overview [1]

In Figure 2.1 the general overview of the different levels of topology is provided. How these topologies interact is very important. Broadcast video will typically be sent through different topologies by use of multicast. This provides an efficient way to distribute video broadcast. Another method of broadcasting video is the use of a Real Time Protocol (RTP) in the transport layer. RTP allows for error correction and it also incorporates packet sequencing and time stamps which should be implemented in the next generation of broadcast encoders. Typically multicast video sources are located at the head end. The use of an MPEG encoder encapsulates the video sources into an IP that is assigned a unique group address for multicast distribution. The STB at the user's location will issue an Internet Group Management Protocol (IGMP) join request. This request will be forwarded to create a SPT based on the (S,G) state. Once the tree is established it will provide a pathway to stream the multicast packets down to the end users [1].

2.2.1 Bandwidth Conversation Introduction

With the increase practice of over provisioning networks Internet Network Providers (INP) must find new methods in reducing system stress. The multicast protocol when used in the correct situation can provide bandwidth reduction throughout the network. This multicast protocol is based on RFP and the unicast routing table. The problem associated with such structure is that it does not provide a necessary control scheme to route the multicast traffic in a more efficient manner [14]. Traditional methods to optimize the multicast routing were based on a Steiner tree model. These methods were used in the attempt to reduce the overall bandwidth of the system, but there are certain limitations. The limitations are the result of the use of the Steiner tree. INPs typically are unwilling to share route specifications and work with other INPs to route the traffic more efficiently. INPs are usually connected to their business competitors and such information sharing and cooperation would remove their competitive edge. The type of information that would be shared would in some cases violate privacy and system integrity [3]. Since inter-domain bandwidth conservation is not a viable option the next logical optimization step would be intra-domain routing. The intra-domain network is under complete control of the INP which would allow for any optimization schemes to be deployed rapidly. By changing the routing within the network an optimized path can be formulated which can reduce intra-domain bandwidth consumption. Another problematic area is between the inter-domain and the intra-domain networks. These two different networks can be seen in figure 2.2. Essentially these two domains represent two different topologies. One example of two different domains would be an internet service provide and a large organization. Multiple connections between these domains would be considered as a multihoming topology. Typically there is high congestion between these two points [4]. As the cost of high speed links decreases more INP are investing in a multi-homed topologies. This topology creates multiple links between the intra and inter domain. This serves two purposes; one for load balancing and the second for failover. Multi-homed topologies could be used in reduction of bandwidth and alleviate congestion on other links. An example of a multi-homed connection can be seen below in figure 2.2 R

represents the intra-domain network which the multi-homed connections are connected to. Essential R domain will contain receiver's r which will represent the end user. At the top of the figure in the S domain is s which represents a multicast source which will propagate down through different domains. The source domain has an aggregated IP address P_s. N in the figure represents adjacent domains which are connected to the R domain by border routers $b_1 \ldots b_n$[3]. These border routers provide a connection between the domain R and the other adjacent domains N which provides the multi-homed connection.

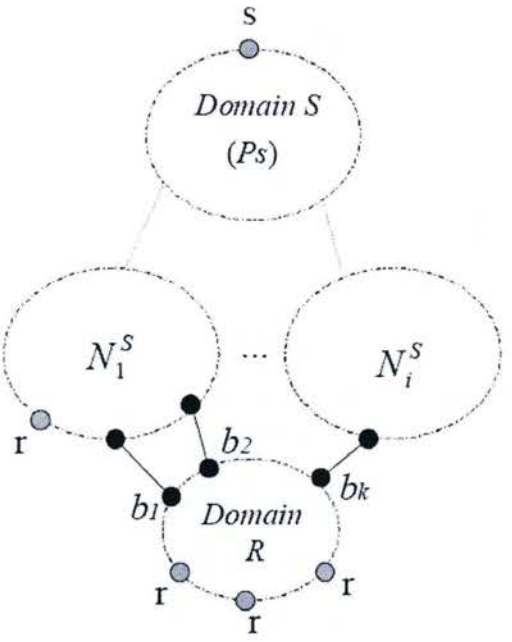

Figure 2.2: Multi-homed Connection [3]

This overview of a typical network topology has been simplified to demonstrate the structure that must be considered when optimization is considered. Since it has already been established that inter-domain optimization is not practical, the task now is to consider how to optimize intra-domain structure such as domain R. There are two main aspects which have already been described. The first aspect is which border router to select, to minimize bandwidth consumption. In conjunction with the first aspect is the second, which is how to route the multicast stream within domain R to reduce

consumption [3]. These two different optimization tasks can be correlated to further reduce bandwidth consumption.

2.2.2 Implementation of Bandwidth Conservation

Traditionally a specific algorithm would optimize a path within a topology and then it would be implemented using MPLS. MPLS can be used to control a multitude of different traffic depending upon the network operator's specifications. While implementing an MPLS path is one way to control the multicast stream, there are scalability limitations. Within the intra domain network a multi-topology protocol could be deployed. This would avoid the use of using MPLS paths which would be required to route the multicast traffic. This multi topology (MT) extension can be applied to IGP [5,6] within the intra domain topology. This extension allows for is the use of applying different link weights for different protocols. A link weight could be applied for all unicast traffic, while another weight could be applied for multicast traffic. This is the main premise of this paper which is how to effectively control the multicast stream to provide optimal bandwidth consumption. This intra domain routing is based on IGP link weights. A comparative analysis has been conducted and shows that using multi topology to control link weights over IGP is an effective way of controlling traffic compared its MPLS counterpart [7,8,9]. This type of traffic flow is considered to be Traffic Engineering (TE). It has been shown that intra-domain link weights is an effective means in controlling traffic with legacy routers [10] and by applying optimization techniques to this method bandwidth conservation can be realized.

2.2.3 Multi-Topology

The use of Multi-Topology is used as an extension of the existing IS-IS and OSPF protocols. It allows for the ability to define link weights for each link based on the type of application that is required. For example when using the Multi-Topology OSPF (MT-OSPF), the MT Identifier (MT-ID) bit with a value of 1 will signify that MT-OSPF is

explicitly used for multicast. The advantage of Multi-Topology when used in conjunction with Multicast IGP M-IGP is that the INP can specify link weights just for the M-IGP. This change will not affect any of the other protocols. Multi-Topology in this section will be explained in detail. This is to allow a greater understanding of how it is used and implemented. M-IGP will handle the intra domain routing aspects. The multicast BGP domain (M-BGP), contains certain fields to identify Address Family Information (AFI) and Sub Address Family Information (SAFI) during BGP routing updates. Essentially this information allows for identification of different traffic flows. When the AFI = 1 and the SAFI = 2 signify that this BGP message will contain IPv4 multicast group [3]. By modifying the group this can be adapted to the discussed optimization model. This will allow for link weight assignment to be placed. The figure 2.3 shows the optimization of M-IGP and M-BGP.

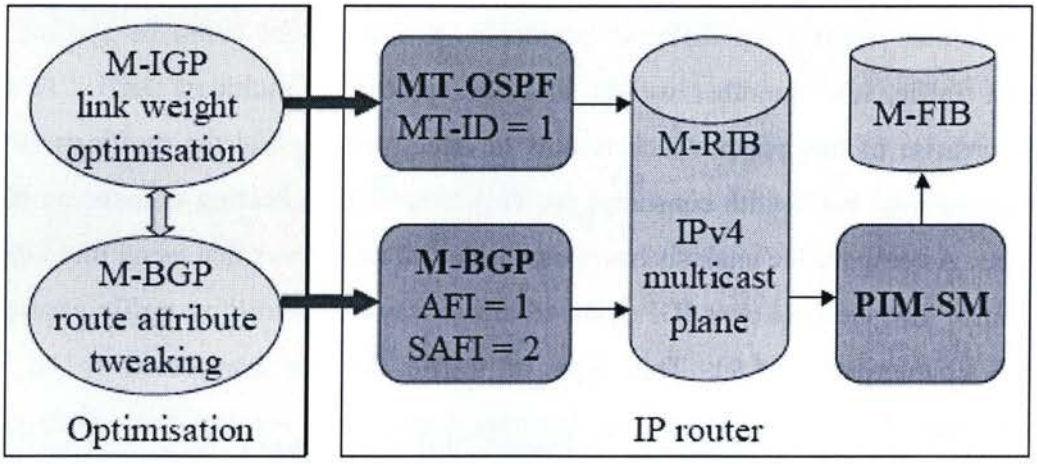

Figure 2.3: Multicast Traffic overview [3]

In figure 2.3 an optimization algorithm would be run to compute optimal link path for Both M-IGP and M-BGP. They are correlated to provide maximum bandwidth conservation. Once the optimization has been complete the MT-ID tag is used to define the link path for M-IGP. For M-BGP fields, such as AFI and SAFI, are alter to indicate a particular link weight structure. These two pieces of information are then correlated within the IP router under Multicast Routing table M-RIB for each source prefix. The

multicast forwarding information base (M-FIB) contains incoming interfaces and outgoing interfaces for each group [3]. This allows routing control of multiple groups that have been optimized within the network.

2.2.7 Inter & Intra-Domain Link Optimization

The need to reduce bandwidth consumption of a multicast topology is becoming more prevalent within the research community. The concepts that were developed in [3] provide important ground work in the field of multicast optimization. For this reason this thesis will focus on these concepts and use paper [3] as a benchmark for comparing the proposed SRP algorithm which will be discussed in further detail in Chapter 3. The algorithm developed in [3] will be referred to as Border router Rendezvous Point (BRP) this is attributed to the characteristics of the algorithm.

BRP uses a novel approach in link optimization, by modeling the source and the receivers set as a Steiner tree. This approach is further expanded to incorporate multiple border routers within the inter domain environment. The time complexity to solve a Steiner tree approach for a large node set is not feasible. The BRP algorithm uses the genetic algorithm (GA) to create a heuristic approach when modeling the Steiner tree. Figure 2.4 outlines the overview of the BRP algorithm. The BRP algorithm optimizes multiple multicast streams within the intra-domain environment. First the genetic algorithm defines the initial routers sets. The algorithm proceeds to group all multicast traffic based on IP prefixes. These prefixes are then sorted based on bandwidth consumption. The prefixes are selected for optimization based on their bandwidth consumption. Optimization is applied in the form of modeling a Steiner tree. Finally a check is used to determine if the optimized path does not consume more bandwidth than what is available. This process will be repeated until all streams have been optimized.

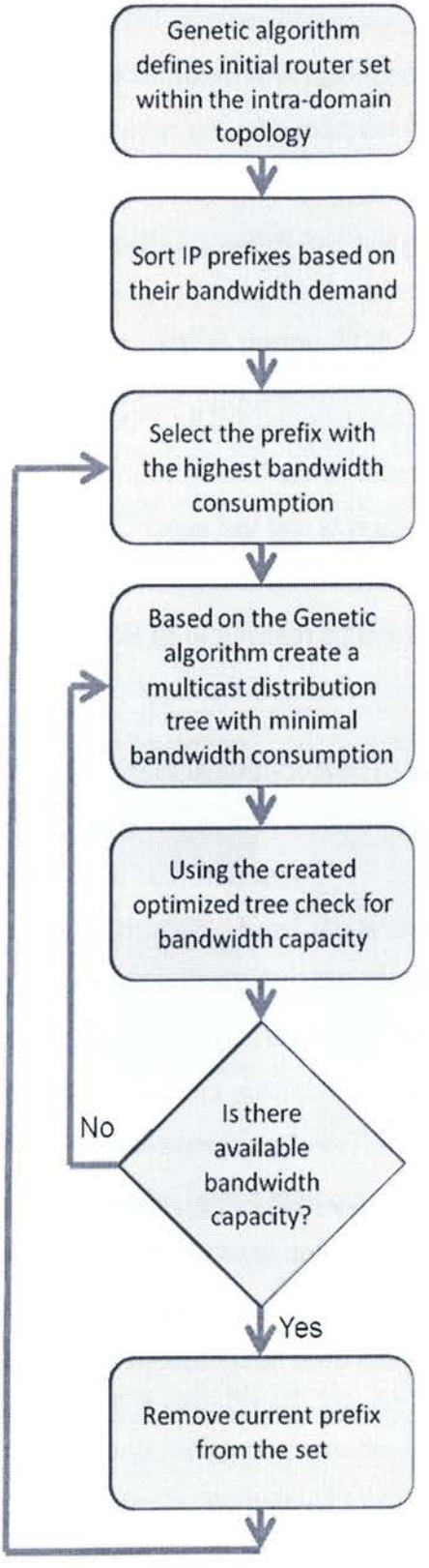

Figure 2.4: BRP Overview

The effectiveness of the BRP algorithm, compared to other forms of optimization, is demonstrated in figures 2.5 and 2.6. The BRP algorithm has been defined as C-HPR-GA within these figures. The topology used for testing of these figures was based on GEANT [11] network which consists of 23 nodes and 76 unidirectional links. The BRP algorithm in this thesis is the C-HPR-GA algorithm in [3]. Figures 2.5 and 2.6 represents an intra-domain topology, the overview of the C-HPR-GA algorithm (BRP) can be seen in the previous figure 2.4. It was assumed in the topology that each prefix can be reached by a maximum of 50% of border routers within the intra-domain topology. Each links bandwidth is scaled to 10^4 units. When using the genetic algorithm crossover, mutation values have been varied between 0.3 and 0.001. Figure 2.5 represents a benchmark comparison between different algorithms. One of the differences between the algorithms is to use a Hop Count (HC) approach instead of using the genetic algorithm to compute the optimized path. These two different optimization algorithms are further subdivided into an uncontrolled Hot Potato Routing (HPR) algorithm U-HPR-HC or a controlled HPR algorithm C-HPR-HC. The HPR approach is a buffer less design for routing. In HPR, single packets will be sent one at a time to the destination. The last algorithm Single-GA is similar to C-HPR-GA (BRP) algorithm except it limits its optimization to one border router. It can be seen in Figure 2.5 that the bandwidth conservation ratio is greatest for U-HPR-HC which has a ratio of 78%. This indicates it conserves 28% of the intra domain bandwidth resources. The C-HPR-GA conserves 15% of intra domain resources. In comparison, intra-domain conservation of C-HPR-GA is not as efficient as U-HPR-HC. C-HPR-GA compensates by load balancing on inter domain links. This can be seen in Figure 2.6, where C-HPR-GA provides the lowest link utilization compared to the other algorithms [3]. This tradeoff provides the best inter and intra-domain optimization.

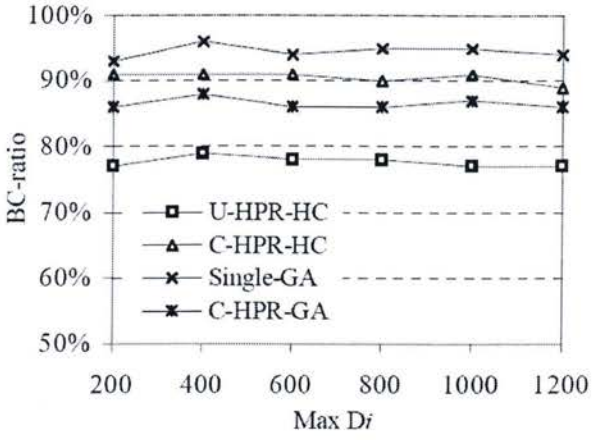

Figure 2.5: Comparative BC-ratio where $\alpha = 10^3$ [3]

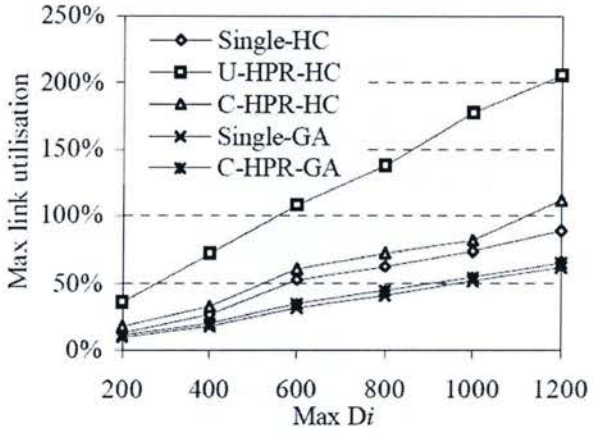

Figure 2.6: Inter-domain link utilization where $\alpha = 10^3$ [3]

While C-HPR-GA (BRP) does provide optimization in both inter and intra-domain bandwidth conservation it does increase the latency of the packet. This type of optimization is concerned with bandwidth consumption and not in latency. Another issue is the number of reiteration of the genetic algorithm. The computational time will greatly increase based on the size of the network. While the genetic algorithm is used to reduce this computational time it still remains based on [3]'s setup. They have chosen to compute all of the optimization algorithms remotely. While this does reduce the computational time, it does not address the situation where topology changes will require

constant updates because of potential of greater external system resources. Another situation is link failure which has not been addressed in this paper. If a failure were to occur it could compromise the stream. Since this type of optimization does not consider link failure, there would be a longer delay in stream recovery.

2.3.1 Fault Anomaly Detection

The described technologies in the previous sections all provide the foundation for providing services and tools so that the end user will benefit from IPTV. The reliable distribution of such applications is the main goal of this paper. What has been described so far is the main distribution and optimization of such a network topologies. While optimization will create an efficient network within an intra-domain environment it does not guarantee reliability. There are a multitude of different faults that can disrupt service. While it may not be possible to completely predict these faults, certain measures can be implemented to source and possibly create a topology or algorithm to reduce the recovery time. The research in [3] has acknowledged that more work needs to be focused on optimization and fault interaction. By exploring different faults it is possible to analyze their structure so that they can be dealt with.

The topologies used for IPTV and similar applications are quite vast. In section 2.1 typical distribution of IPTV multicast can occur within any topology and across any network. For this reason in depth analysis is required to fully understand the inner workings of possible anomalies. The BGP protocol uses Autonomous System (AS) which is a collection of IP networks and other resources. This information is used to dictate routing and maintain link information.

A BGP router will periodically send routing announcements containing messages of prefixes regarding its ASs or of ASs that it is linked to. These routing messages that are used to establish traffic flow are vital in ensuring the correct traffic pattern. Problems can

arise if a misconfiguration occurs. This can cause detrimental effects and create massive slowdowns throughout the internet [12].

One such misconfiguration is route hijacking. This occurs when a BGP router advertises a route that it has no access to. Fundamentally what occurs is that packets will be dropped. Another type of misconfiguration is route leakage. This occurs when one BGP router sends more routes to a peer than it is capable of handling. When this happens, the peer router will be overloaded and will affect the stability of the routing. This effect can also trigger route oscillations which also consumes the recourses of the router [12].

Traffic anomaly detection is typically based on a statistical approach. This is achieved by comparing current statistical information regarding the routing with previous history. By this comparison if the current statistical information deviates from historical information then it will be assumed a possible anomaly is present. When BGP updates occurs, the system will monitor the frequency of these updates as well as the time that it takes for a prefix to converge. By basing anomaly detection on historical information it is easier to implement a detection scheme for anomalies. Another advantage of this type of statistical detection method is that it is extremely efficient when processing massive amounts of data. It does have short comings in its ability to detect complex problem. Another major disadvantage is that it requires fine tuning in order to set a threshold value. The threshold value will be different based on the type of network in use. If this threshold value is set to high or low it can cause significant false positives or, even worse case it does not detect certain anomalies. The threshold value has been defined as a "magic-number" in [13] which needs to be adjusted for proper operation of the algorithm. A learning based approach as described in [13] has a few key differences. Both a statistical and a learning based approach use history as a comparison. But the learning based approach compiles its history differently then a standard statistical approach. It also has different testing methods. In essence how this is achieved is that BGP update behavior is represented as a vector. This vector contains certain aspects of the BGP updates and can be used to map certain patterns in a multidimensional vector space. Based upon these mapped points, if a

point is mapped far away from the initial point of the domain, it is considered a possible anomaly. If another point is mapped to another location deemed as a normal location this would signify normal operation. This type of research is headed towards identifying clusters of normal operation. This will allow for proper tagging of anomalies. This has different advantages over other approaches such as dealing with a variable number which identifies the detection rate. The other advantage is that this approach incorporates a shift-invariant property. What this means is if a burst of updates occurs randomly in time it shouldn't matter when it occurs but how uniform the burst is. This uniformity is in regards to the timing of the updates within the burst. This method is achieved by using wavelet transforms which is a signal processing technique. Essentially it maps signals into time-frequency domains so that it can be easily represented [13]. This is different then using the PCA method to compare the statistical correlation between each BGP update. With the PCA method if a deviation occurs from the predetermined path it is considered an anomaly.

2.3.2 Data collection

Routing information was collected from Border Gateway Protocol (BGP) and Interior Gateway Protocol (IGP) routing messages. This collection was accomplished by using Packet Design Route Explore (REX). REX is used to capture all internal routing information between BGP routers. This information does not consider route withdrawal attributes. For this reason routers that are connected with REX will relay their full routing information including all of their attributes. REX also monitors adjacency IGP routers and collects all link stats information that may be developed from these links [12].

Paper [12] results are based on two different data sets. The first data set was collected at U.C. Berkeley on August to December 2003. U.C. Berkeley consists of a four-area OSPF which is interfaced with IGP. REX was initiated between the BGP edge routers. This process was then repeated to an ISP. The statistics that REX collected is seen in table 2.1 below.

	U.C. Berkeley	US
BGP Nexthops:	13	9150
Prefixes:	12 600	200 000
Routes:	23 000	1 500 000

Table 2.1: Statistical Data collected from two different Sources [12]

2.3.3 Data Analysis the PCA approach

The data was analyzed using two different methods; an analytical model and a visual model. The visual model that was used is called TAMP (Threshold and Merge Prefixes). The analytical model that was used was called stemming. TAMP takes into consideration all the prefixes. It then creates a virtual tree based on these prefixes. Weights are then assigned to each edge of the tree based on the rarity of the prefixes that are implemented on a particular edge. TAMP will then proceed to prune all nodes and edges that represent less then 5% of the entire graph. TAMP only represents a snapshot in time.

TAMP can be used to identify certain types of misconfigurations such as backdoor routes and load balancing situations. It can detect when a back door route occurs such as in Figure 2.7. The backdoor route is defined between 128.32.1.222 and 169.229.0.157.

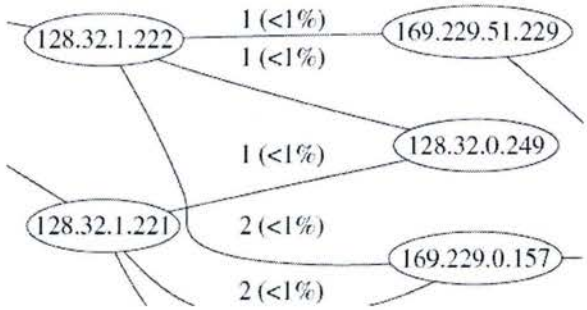

Figure 2.7: TAMP Visualization [12]

Stemming is the analytical model which correlates the repetition of an AS path to systematically detect an anomaly. The method of stemming is based on the principal component analysis model. By monitoring BGP route withdrawals, a sequence can be extrapolated. AS has been defined as a certain sequence such as $a_1..a_n$. These withdrawals represent a certain sequence which can defined as $c = xha_1...a_np$. Where x is a peer that withdrawals from a certain prefix p. These sequences occur over a period of time and can be defined as a stream; where $C= c_1..c_2..c_m$ represents the stream. The algorithm will then tabulate pairs of adjacent ASs that occur in the C stream. By ranking these pairs the algorithm can determine potential problem areas. An example of this can be seen below in figure 2.8. The adjacent pair 11423-209 occurs multiple times which indicates a problem location. This is defined as the failure location.

```
W 192.96.10.0/24 11423 209 701 1299 5713
W 207.191.23.0/24 11423 11422 209 4519
W 192.96.10.0/24 11423 209 701 1299 5713
W 212.22.132.0/23 11423 209 1239 3228 21408
W 203.14.156.0/24 11423 209 701 705
W 209.5.188.0/24 11423 11422 209 1239 3602
W 12.2.41.0/24 11423 209 7018 13606
W 12.96.77.0/24 11423 209 7018 13606
W 62.80.64.0/20 11423 209 1239 5400 15410
W 62.80.64.0/20 11423 209 1239 5400 15410
```

Figure 2.8: Prefix Withdraws [12]

In conjunction with monitoring correlations, stemming can use TAMP as a visualization to give a more comprehensive understanding of what is transpiring. Stemming can be used to detect a single route oscillation because of how it correlates to the data. If stemming is left to run over a short period of time, after a few hours these oscillations will appear as a strong correlation and can be easily detected. Stemming can also detect misconfigurations regarding route leakages, where routes are directed to a longer path, which may not be desirable. This depends on the policies that an ISP might have in place.

The next step that [12] is trying to address is how to correlate routing policies defined in the routers configuration file with the actual routing that is taking place. Routing policies are not announced in AS events and this makes monitoring them more difficult. Stemming uses strong correlation to identify problems but, without cross comparison with a configuration file, it is difficult to analyze the situation to provide the best course of action.

Ongoing work is being done to incorporate traffic into the analysis to detect routing misconfiguration which could distribute prefixes based on traffic load. This presents a problem when a small block of prefixes may contain 90% of the traffic load. By rearranging the prefixes it would provide a more balanced routing of the traffic. Traffic and routing are both interrelated and both need to be addressed simultaneously in this situation.

2.3.4 The use of Wavelets

This section explores other techniques to identify potential network anomalies. Understanding anomalies and their occurrences indicates the importance in a failure recovery model. Wavelets are used in the process to transform prefix updates into a vector representation. A data set is defined in 24 hour intervals. This represents one update in the actual calculation. This value is adjustable and can change depending upon the detection model. Every update i is modeled in a sequence of S and has a length of n. A discretized continuous wavelet transform is performed on this sequence. The Haar version of wavelet transform is used. This wavelet is described below in equation 2.1. Where τ represents the translation and δ represents the scale. Now the discretized transformation is defined in equation 2.2 [13].

$$\Psi(\frac{x-\tau}{\delta}) \quad \textbf{...equation 2.1 [13]}$$

$$\gamma(\delta,\tau)=\sum_x S(x)\cdot\frac{1}{\sqrt{\delta}}\Psi^*(\frac{x-\tau}{\delta})$$ **...equation 2.2 [13]**

This transformation has a set of scales from δ_0 to δ_τ. For the initial wavelet a scale of 20 seconds has been chosen. This means $\delta_0 = 20$ and scales thereafter are $\delta_{i+1}=2\delta_i$. τ can take on values from the set {1,2,...,n) and this transformation results in $\gamma(\delta,\tau)$, which is τ for time and $1/\delta$ for frequency. After this transformation the original data set will be even larger. This poses a problem in that the solution is to only consider large burst values of t for. $\gamma(\delta,\tau)$. By taking the peak values of the bursts and maintaining the duration between other peak values, this allows for a reduction of data in the transformation set. Another technique for reducing the data is to take approximate values of the peaks of $\gamma(\delta,\tau)$. This technique detects the largest value from $\gamma(\delta,\tau)$ and sets it as γ_{max}. The interval for these peak values will be between (0, γ_{max}]. The data is then stored in a histogram. Using this histogram and other properties of the histogram the bin will be $(0,v]$. This will be defined as: $(v(1+\varepsilon)^i, v(1+\varepsilon)^{i+1}], i=0,1,2....$ The default experimental values that where initially used provided excellent results and are detailed as follows v = 0.1 and ε = 0.5. Essentially a histogram is a way of storing the peak values in a form of a vector. This is a more efficient way of storing and accessing the required data. Detailed information can be found in [13].

Described above was how the actual data was compressed and stored. Each different method requires some sort of database to index the available collected data. This is required regardless of the detection method that is used. The processing of data for this learning based approach is to cluster the data. By clustering this data, trends can be detected. If the majority of the routes behaves a certain way, this will show up as a large cluster and anything deviating from it represent a possible anomaly. Clustering is supposed to increase efficiency. First vectors have been created by using a histogram then clustering is applied to the prefixes. Figure 2.9 shows cluster sizes. From this figure it can be easily determined that the majority of prefixes are in one cluster grouping.

There are approximately 170 prefixes in the first cluster. There are none in the second cluster and so on until the 7^{th} cluster is reached. This is where possible anomalies are clustered. In the next figure 2.10 from a-d different values of k are varied. This k values represents the number of clusters in a data sample. By adjusting the size of k it will change the sensitivity of the detection algorithm. The size of a cluster will determine the granularity of the detection. The value of k determines the number of categories which can increase or decrease the detail of the cluster. The negative aspect to this is that computation time will be increased. This may not be significant over a small sample however it will present a problem over larger samples. The smaller the k value the less storage space is required for the computation. In figure 2.10 (a) it can be seen that there are over 150 points in the cluster. This is using the cluster grouping of five. This gives a rough estimate of potential anomalies. By increasing this value from 5 to 30 a distinct pattern can be seen. These graphs represent BGP updates and this model can also be adapted for other types such as OSPF updates. The same procedure would be duplicated and any variance from the expected values would be considered an anomaly.[13]

The points in the cluster share some commonality for a particular prefix. Work has been done on correlating multiple prefix clusters. This allows for the ability to examine the update relationship between prefixes and to detect possible anomalies using the same detection clustering methods. This work is in its preliminary stages. It still requires more fine tuning to isolate anomalies.[13]

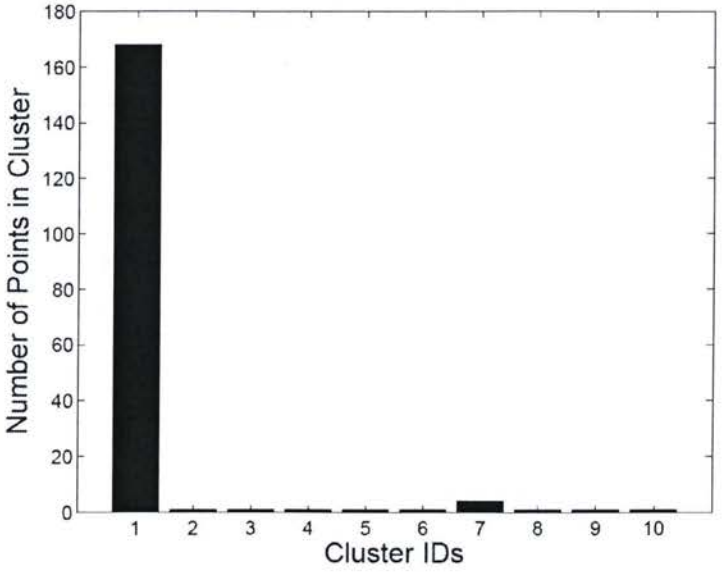

Figure 2.9: Cluster Formations [13]

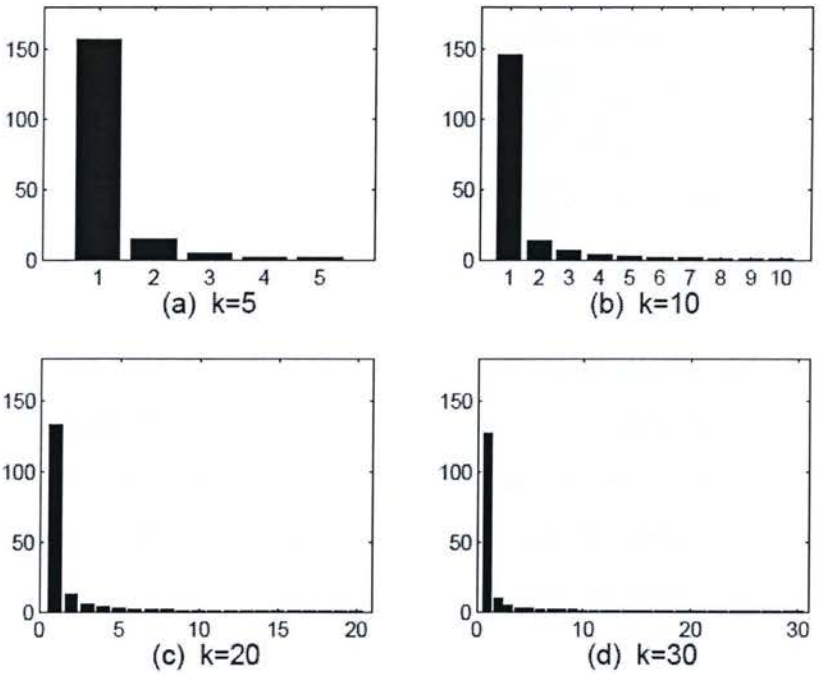

Figure 2.10: Cluster Formations with varying k values [13]

2.4.1 Traffic Anomaly Detection

Traffic anomaly detection is related to router anomaly detection. These two detections models are interrelated and should be addressed concurrently. Most of the work currently available does not address both of these issues. This section will address some key issues and give an overview of some general concepts. By introducing these concepts, a greater sense will be gained about the statistical analysis. In paper [12] it was discussed how a detection scheme was based on route withdrawals as it was discussed that future research would incorporate traffic. The next progressive step is to correlate route withdrawals with traffic.

The reason traffic is important is because traffic is routed based on how the router is configured. By using this method of detecting how the traffic is routed a model can be generated to detect possible errors. A traffic model can be used to correct routing anomalies in load balancing. Some routes may be configured incorrectly and more traffic may pass through these routes. This can be a routing configuration issue and needs to be addressed. This issue has been identified in [12] and current resolutions are being explored.

As with analyzing routing anomalies, traffic anomalies involve the collection of data followed by the processing of that data using a variety of different methods. Analyzing this data can indicate if an anomaly is a malicious attack, large file transfer or, more importantly a router failure or another type of misconfiguration. In paper [16] they have also used a wavelet model to detect anomalies.

2.4.2 Using PCA for traffic anomalies

The first step in applying the PCA algorithm is to define a traffic matrix. This traffic matrix can hold valuable information in the form $v_{i,j}$. The matrix can represent packets or bytes over a period of time and can be used to organize a multitude of different information. The way the traffic matrix is structured is that one dimension of the traffic matrix will contain information regarding the time whereas the other dimension will pertain to actual data such as bytes information or entropy of addresses in relation to time. As discussed in paper [15] m is a column in the traffic matrix. The columns represent IP flows. When PCA is applied to the matrix, it will produce orthonormal vectors. These vectors represent the highest variance when compared to the original matrix. They belong to the k subspace where k <= m. This k will refer to a normal subspace of the network. The method in which traffic anomalies are detected is when k is removed from the subset; where n-k represents the removal of the normal subset, which would leave a subset which contains anomalies. They describe this process in terms of random variables which are in the form of a m x n matrix. The only consider cases when m < n which is defined as greater observations than variables. This procedure is just used to model the actual flows. The traffic matrix is a representation of vectors in relation to time. This is defined as $\vec{v}$. Using the model which has been defined above then projecting $\vec{v}$ will show which region this vector is in. This will indicate whether it is in an anomaly subset. [15] The steps for applying PCA to IP flows are as follows. These steps have been outline in the figure 2.11 as well.

1. Data logging: This was achieved by reading Zebra data logs of all BGP messages. These messages contained egress and prefix pairs. By combining this information and parsing the log files, a complete detailed path can be formulated.

2. Traffic aggregation is applied to divide the data into different categories. By grouping certain characteristics it greatly improved the success rate of PCA detection model.

3. The next step is to apply entropy timeseries function to the data set. This data is represented by $\vec{v}_i$ vector. It will contain four distinct data entries which will be in the form of $v_{i,j},\ldots,v_{i,j+3}$. These entries are defined as source IP, destination IP, source port and the destination port. Entropy will be applied to the vector. Since the vectors behave like random variables, entropy can be applied. An example of how this applies is as follows. When a lot of traffic is received on a server port 80, the entropy of this port will decrease towards 0. This indicates that the probability of more traffic connecting to port 80 is increasing.

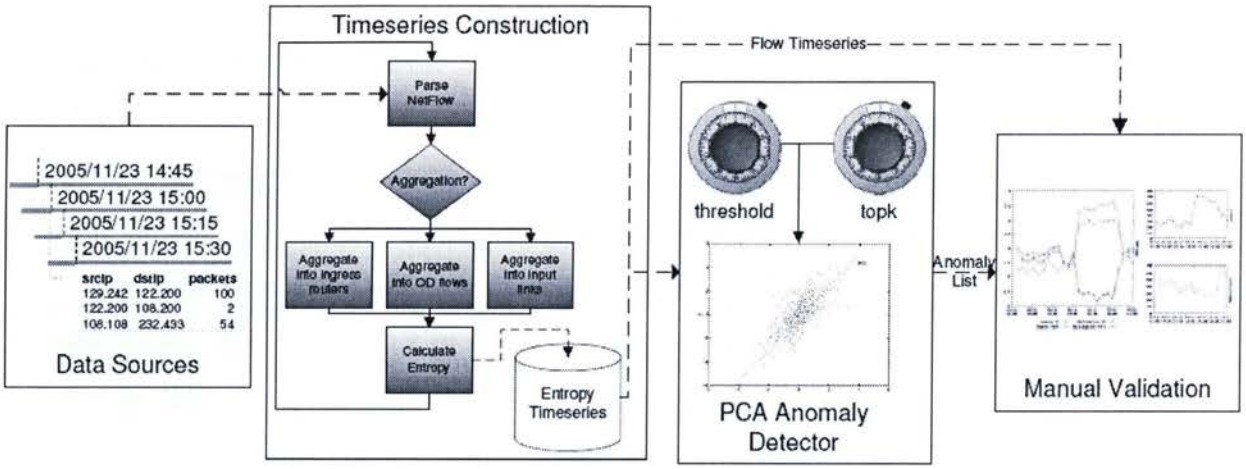

Figure 2.11: Step Overview [15]

Manual validation is required to check the results of the PCA algorithm. This will allow for identification of false positives. Another advantage of manually validating the results is that PCA may not flag precise moments in time. When manually validating, subtle changes can be detected and may be considered as errors. Depicted in the figure above are two variables in the PCA anomaly detector. These are fine tuning variables which are required to change the sensitivity of the PCA algorithm. Tests which have conducted have shown that this false positive rate can vary from 3% to 16% [15].

2.4.3 Analysis

By monitoring this traffic, based on source and destination IPs, the path that will be taken can be formulated. This can be accomplished by including which routes are required to achieve this path. By knowing the next hop in the path for the traffic, routing anomalies can be detected. They will show up as inconsistencies. If QoS requires a certain path and it deviates from the path, this will show up in the PCA analysis. Port information can also indicate excessive routing problems where an increase of traffic to particular addresses may be caused by routing flaws. This traffic can be applied to a particular protocol and correlated with other routing information to indicate if a potential problem exists.

2.5 Background Overview

Each method in the subsections above describes anomaly detection and IPTV distribution over a multicast network. Chapter 3 correlates these two ideas by expanding on section 2.2.1 to include failure models. Failures can be result of different anomalies such as route oscillation. These anomalies typically manifest as either a failed link or reduced capacity of a link. This background information provides insight on failures so that an optimized solution can be developed to reduce the failures effect on the multicast topology.

Chapter 3: Joint Optimization with failure recovery

3.1 Introduction

This section addresses optimization of packet routing for inter and intra domain routing by applying the SRP algorithm. The SRP algorithm optimizes both load balancing and bandwidth conservation within a topology across two border routers. This is achieved by correlating different optimization instances and problems described in the background, chapter 2, which are referred to as BRP (border rendezvous point). By understanding how link failures occur and the types of failures, optimization can be applied to reduce their effects. BRP provides valuable information regarding link integrity. Anomaly detection can be an invaluable tool that can be used to discover and increase the optimization of inter and intra domain link optimization. The approach explored in this section is to quantify the effects of a link failure and determine how catastrophic such a failure would be. Two different models will be explored. In [3] they have explored just optimization based on bandwidth conservation. This idea has been further expanded to include border router link failure. This is then compared against a new approach of using a SRP (single rendezvous point). The SRP will provide stability within the topology for all nodes below it. This stability is achieved by exploiting the functionality of the RP (rendezvous point). If one border router fails in a multihomed topology only the PIM Join message from the RP to the second border router is required to reestablish the connection. The multicast tree below the RP remains unaffected. The placement of the RP is critical to ensure no disruption of service. Both the RP and BRP models are described below outlining their characteristics and implementation.

3.2 Bandwidth Conservation and Load Balancing

Link optimization to minimize the number of links in a multicast distribution tree is a key aspect in bandwidth conservation. In figure 3.1 two different paths are defined, where the links in the paths are represented as dark arrows. Both border routers b_1 and b_2 have access to the source. The total number of links used in this routing scenario is 6. Since both border routers have access to the same source, their locality within the topology can be used to create an optimal routing path. In version (b) the number of links to satisfy all the receivers is reduced to 3 [3]. This simple example shows the benefit of link optimization.

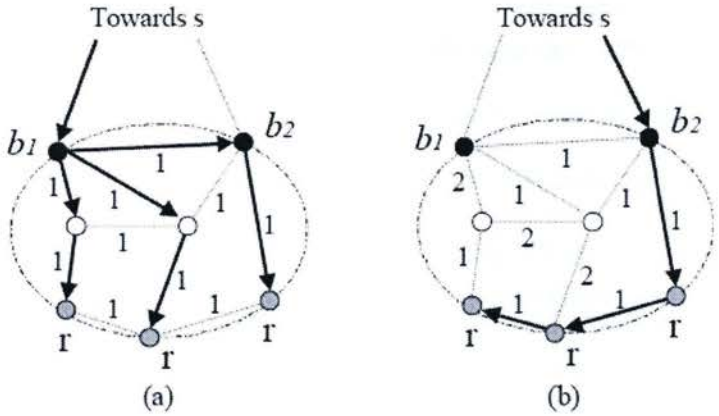

Figure 3.1: Link Optimization [3]

3.2.2 Intra-Domain Cost model

The cost model for the intra domain routing is given in equation 3.1 to 3.2. These equations serve two distinct purposes; intra-domain bandwidth conservation and inter-domain load balancing [3]. The network topology can be written in a graph notation G=(V,E). Where V denotes the router set in the topology. E denotes the physical link

between the set of routers. Certain characteristics including bandwidth limitation of a link are calculated in the equations below. Nodes of the set V are categorized into two distinct groups; where V_A represents the set of access routers and V_B represents the set of border routers. The border routers of V_B are assumed to be connected by means of an inter-domain link with a bandwidth capacity of C_b.

There is a set of prefixes $P = \{P_1, .., P_k\}$ which are reachable through all the border routers in V_B. There are also t multicast groups, $m_1, ..., m_t$, receivers are connected to the access routers that has been defined as Va. Access routers will carry the multicast stream and during simulation they are also considered as receivers. The source address $s_i (0 < i < t)$ for a multicast group is assumed to belong to a prefix $P_j (0 < j < k)$. Each receiver set R_i receive packets for its corresponding multicast group m_i. The receivers in the simulation are considered as a set of access routers $R_i \subseteq V_A$. Each access router is connected to the multicast tree T_i with a bandwidth demand of D_i. Load balancing across border routers is achieved by selecting each prefix $P_j (0 < j < k)$ and assigning this prefix P_j to an M-ASBR b_j router for its ingress point. The next step is to assign link weights w_{uv} so that overall cost of the multicast trees can be minimized [3]. This will ensure that the optimized path is followed when the reverse path forwarding (RPF) check is performed.

Equation 3.1 provides formulation of intra-domain bandwidth conservation. Essentially it sums the bandwidth demand D_i for all the links that are contained within T_i. Once this procedure is completed it will provide a cost function for m_i. The external summation calculates the cost of the multicast trees for all the multicast groups $m_i, ..., m_t$. This provides a total cost function for the entire intra-domain network. This calculation needs to be minimized to provide bandwidth conservation. A Steiner tree algorithm is applied to this to minimize the cost function. The problem associated with the Steiner tree is that it

is a NP-hard problem. To resolve these issues the genetic algorithm is applied in conjunction with the Steiner tree [2]. Since the RPF check is performed over the shortest path to the source, it fails on the Steiner tree. To resolve this issues link weights are calculated in such a manner that they create a shortest path tree overlaid on the Steiner tree. This ensures that the RPF check does not fail based on the Steiner tree.

There is a possibility of over provisioning the inter-domain links. To avoid this, equation 3.2 minimizes the maximum link utilization. Where u_b^{inter} represents the link utilization; equation 3.2 is used to minimize the maximum link utilization of the inter domain link.

$$\text{minimize } l^{\text{intra}} = \sum_{i=1}^{t} \sum_{(u,v)} D_i \times x_{uv}^i$$

$$where\ x_{uv}^i = \begin{cases} 1\ if\ (u,v) \in T_i \\ 0\ otherwise \end{cases} \quad \text{.....equation 3.1 [3]}$$

$$\text{minimize } \max(u_b^{\text{inter}} = \frac{\sum_{i=1}^{t} D_i \times y_b^i}{C_b}) \text{ for each } b \in V_B \text{ .....equation 3.2 [3]}$$

$$\text{Where } y_b^i = \begin{cases} 1\ if\ b \text{ is selected as the ingress of m}_i (i.e.\ \text{root of } T_i) \\ 0\ otherwise \end{cases}$$

3.2.3 Inter & Intra-Domain optimization

Inter and intra-domain optimization is based on the principle of load balancing the multicast streams across multiple border routers V_b and across intra domain routers V_a. In equation 3.1 bandwidth optimization is performed in the intra-domain. This cost function measures the total intra-domain bandwidth by including multiple ingress points. In [3] they discuss load balancing across border routers together with the intra-domain bandwidth conservation. This allows optimization of both inter and intra-domain path selection.

The algorithm 3.1 below is the procedure for optimizing both inter and intra-domain bandwidth. This procedure uses a controlled Hot Potato Routing (HPR) algorithm. The HPR algorithm in unicast routing generally provides maximum bandwidth consumption. For multicast application the HPR does not address the issues of packet duplication on each V_a router. For this reason a controlled HPR algorithm is developed. This algorithm is described in algorithm 3.1 which allows for multiple intra-domain routers to share common paths to locate the closest ingress router [3]. The second part of the procedure relies on the genetic algorithm to reduce the possible solution set and find the optimal solution relatively quickly.

The procedure begins by arranging the bandwidth in terms of P_j bandwidth consumption. Once completed P_j is assigned a border router $\bar{b}$ based on certain criteria in the fitness algorithm. This initial step is required and is used as a benchmark for further calculations which can be seen in the while loop. The second part of the algorithm is used to select a second ingress border router. By spanning the newly selected ingress border router with the previous set of border routers, an optimized link cost can be calculated $l_j^{\text{intra}}(B_j \cup \{b'\})$. This step provides load balancing; a control variable λ is used to

provide a threshold for the ingress selection. In [3] the value of 0.5 is chosen for λ. The fitness test in algorithm 3.1 is used to direct the results towards inter and intra-domain bandwidth conservation; where α is a tunable parameter for balancing between inter and intra-domain consumption.[3]

Procedure BRP (C-HPR-GA-Fitness)

Begin

Set the *M-IGP* weight of each intra-domain link in the network according to the chromosome based on the initial genetic algorithm set;

For each prefix P_j

Aggregate group bandwidth demand according to P_j, i.e.,

$$AD_j^{\text{inter}} = \sum_{i=1}^{I} D_i \text{ for } s_i \in P_j;$$

End for;

Sort the prefix list P in descending order according to $AD_j^{\text{inter}} (0 < j < k)$;

For each prefix P_j in the ordered list P

Assign an *M-ASBR* $\bar{b} \in V_B$ reachable to P_j such that

intra-domain bandwidth consumption $l^{\text{intra}}\left(\left\{\bar{b}\right\}\right)$ is

minimized for the groups whose source $s_i \in P_j$ and

M-ASBR $\bar{b}$ has sufficient residual bandwidth for the aggregated demand AD_j^{inter};

Update inter-domain link utilization on $\bar{b}$, i.e.,

$$u_b^{\text{inter}} = u_b^{\text{inter}} + \frac{AD_b^{\text{inter}}}{C_b};$$

$$B_j = \left\{\bar{b}\right\};$$

$|B_j| = 1$; /*Find additional ingresses for P_j */

While $|B_j| < B_m$

Find $b' \in V_B \setminus B_j$ reachable to P_j such that

Intra-domain bandwidth consumption $l_j^{\text{intra}}(B_j \cup \{b'\})$ is minimized and

M-ASBR b' has sufficient residual bandwidth for the aggregated demand AD_j^{inter};

If $l_j^{\text{intra}}(B_j + \{b'\}) < \lambda \times l_j^{\text{intra}}(B_j)$

$B_j = B_j \cup \{b'\}$; $|B_j| = |B_j| + 1$;

Update inter-domain link utilization on b', ie.,

$u_{b'}^{\text{inter}} = u_{b'}^{\text{inter}} + AD_{b'}^{\text{inter}} / C_{b'}$;

End if;

End while;

End for;

$l^{\text{inter}} = \sum_{j=1}^{k} l_j^{\text{inter}}(B_j)$; /* Sum up total intra-domain bandwidth consumption for all prefixes*/

$$fitness = \frac{\alpha}{l^{\text{intra}} + \alpha \times \max(u^{\text{inter}})};$$

End

Algorithm 3.1: BRP Algorithm [3]

3.3 Implementation Issues

The optimization problem formulation and the BRP optimization algorithm presented Section 3.2 is designed to balance the bandwidth load across the border routers together with reducing the total bandwidth consumption of the multicast trees within the domain. Figure 3.2 shows a simple realization of a multi-homing situation for a multicast access network. The border routers in the network perform inter domain load balancing. The figure shows a multi-homing configuration where a multicast network is divided over two different domains. The distribution network is used to distribute a multicast stream such as IPTV to end users in the access networks. The multicast stream is generated at sources which may be part of the distribution network or located in a content network connected with the distribution network. In the figure BR 1 and BR 2 represent two distinct border routers on the edge of the distribution network.

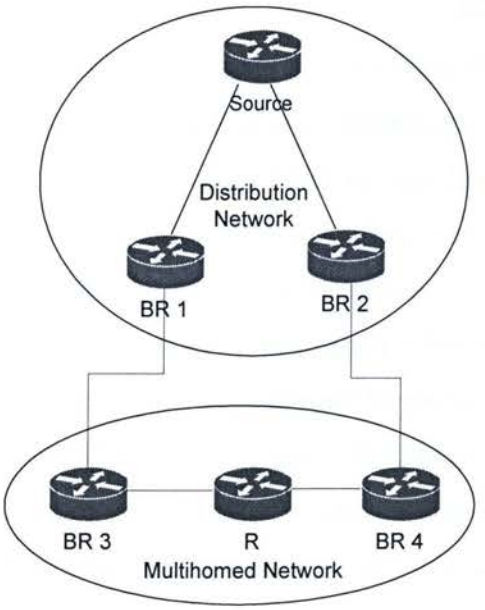

Figure 3.2: Multihomed topology of an access network

While BR1 and BR2 are shown in the same network in this example, typically they are located in two different networks connected with the access network through varying connection type and speed. In the figure BR 3 and BR 4 represent the border routers in

the multihomed access network, whereas R represents an internal router. The widely used multicast protocol which has become the defacto standard lately is PIM-SM (PIM-Sparse Mode) [18]. In PIM-SM a router is selected to be the Rendezvous Point (RP) or the root of the multicast tree. The BRP algorithm assumes border routers to be the RPs, hence we call that as the *BRP algorithm*. The multihomed connection, as shown in figure 3.2, can provide a variety of different services to an organization or ISP. This topology could represent a national ISP with resale to smaller independent ISPs or to an organization. The use of multihomed connections can provide the following services.

- Load Balancing
- Link Redundancy
- Traffic Priority

3.3 The SRP Model

The BRP algorithm in algorithm 3.1 assumes each border router to act as RP for a multicast stream. We call that as BRP model. The problem associated with the BRP model is that if link failure occurs on any border router such as the external link failure or the internal link failure connecting with the network there will be a migration of receivers towards the unaffected border router which acts as other RP. This causes duress within the network topology in terms of the slew of control or join messages propagated through the network along with the delay in resuming traffic flow to the access routers connected with the multicast tree rooted at the affected border router. Further, as links are shifted during a link failure the effectiveness of the link optimization is reduced. The original BRP algorithm calculated bandwidth conservation based on a static topology, this limitation does not allow for a shift in the topology due to a link failure. The BRP model distributes links based on available bandwidth and link utilization. Congestion and high link utilization can form around the border router which can be seen in figure 3.3. The border router br_1 has a high proportion of routers surrounding it r_2, r_3, and r_4. This high proportion of routers surrounding the border router can create an increase of link

congestion. The BRP algorithm performs only checks to avoid over provisioning of the link utilization and not of link density as shown in figure 3.3.

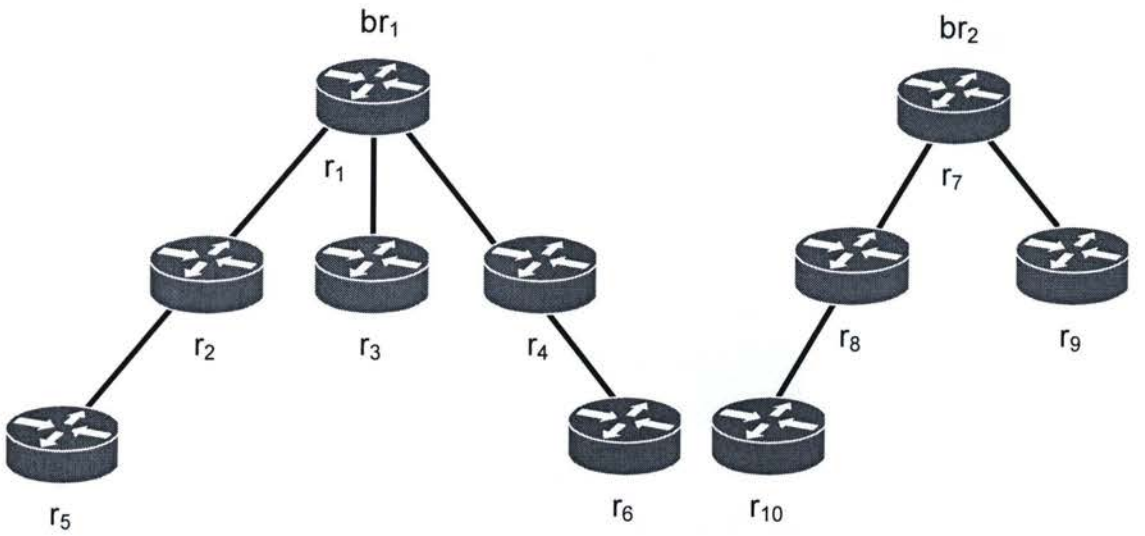

Figure 3.3: Border Router as RP

Figure 3.3 represents one multicast stream that is propagated through br_1 and br_2 to r_2, r_3, r_4, r_8 and r_9; this traffic causes congestion between these routers and decreases available bandwidth for other forms of traffic.

In contrast to BRP, we propose SRP model where any router within the access network can be selected to be the RP. Even multiple RPs can be selected to achieve the benefits of high reliability and availability through redundancy. However, in this thesis we propose and investigate a single RP for a multicast distribution tree. Multiple RPs can be used but for different multicast distribution trees. Since we limit ourselves in this thesis to the recovery from a single link failure, single RP model causes no serious drawback. Further, routers are generally more reliable than links; especially routers hosting RP function are more reliable routers. Our solution has merit as it deals with link failure which is more common than router failures. By creating a SRP router within the topology this high concentration of links can be shifted away from the border router. This reduces the adjacent link utilization of the border routers and reduce the affect on other forms of

traffic. Figure 3.4 shows the SRP model, where only a single link from the border router provides a multicast stream. This is a simplified example but when compared with figure 3.3 shows significant decrease in link utilization of the surrounding br_1 links. This is attributed to shifting the root of the multicast tree to the RP which is located within the intra domain topology. Only one link is required to supply the RP with a multicast stream.

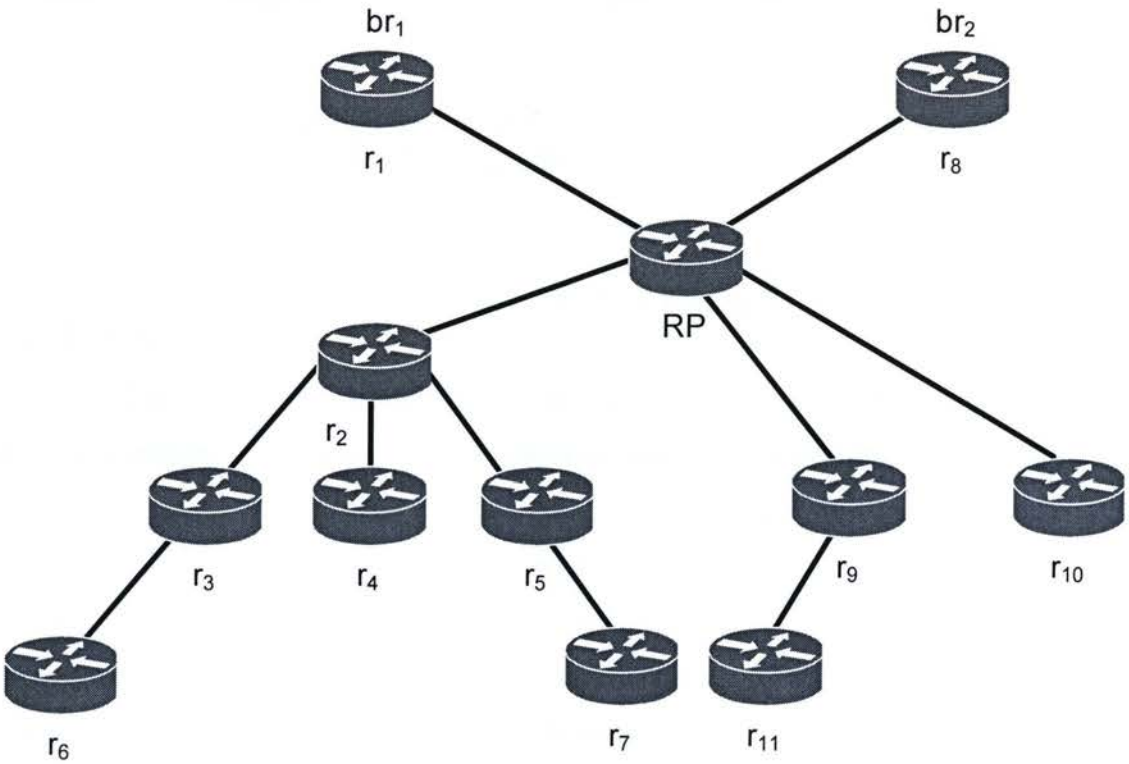

Figure 3.4: SRP model

3.4 SRP Optimization Algorithm

In this section the SRP model is discussed in detail. The main premise is to determine the location of the SRP. The locality of SRP between the border router and the receivers impacts link optimization and system distress during a link failure. Determining the best location for RP in the SRP model is critical because the best placement of RP provides minimal hop distance between each border router and maximum link optimization.

The SRP model relies on the same fundamental principles of bandwidth conservation, through link optimization and load balancing, as does the BRP model. The physical topology T which includes all routers and links is translated into graph theory. Each multicast stream is analyzed to determine the receivers and to select a RP. This information is then used to construct a Steiner Tree. When constructing the Steiner Tree, each link along the optimized path is checked for available bandwidth, to avoid over provisioning of resources. This process is repeated for each multicast group. Figure 3.5 depicts the boundaries of both inter-domain and intra-domain network. The border routers are designated as B1 to B2. The receiving routers are defined as r_1, r_2, r_3, r_4, and r_5. Link optimization described in [3] addresses both inter and intra domain routing of a multicast stream. The SRP model affects both the inter and intra domain routing by controlling which border routers should be selected for the RP. Once this is complete optimization described in [3] is applied to determine the interconnecting routing links between the receivers and the RP.

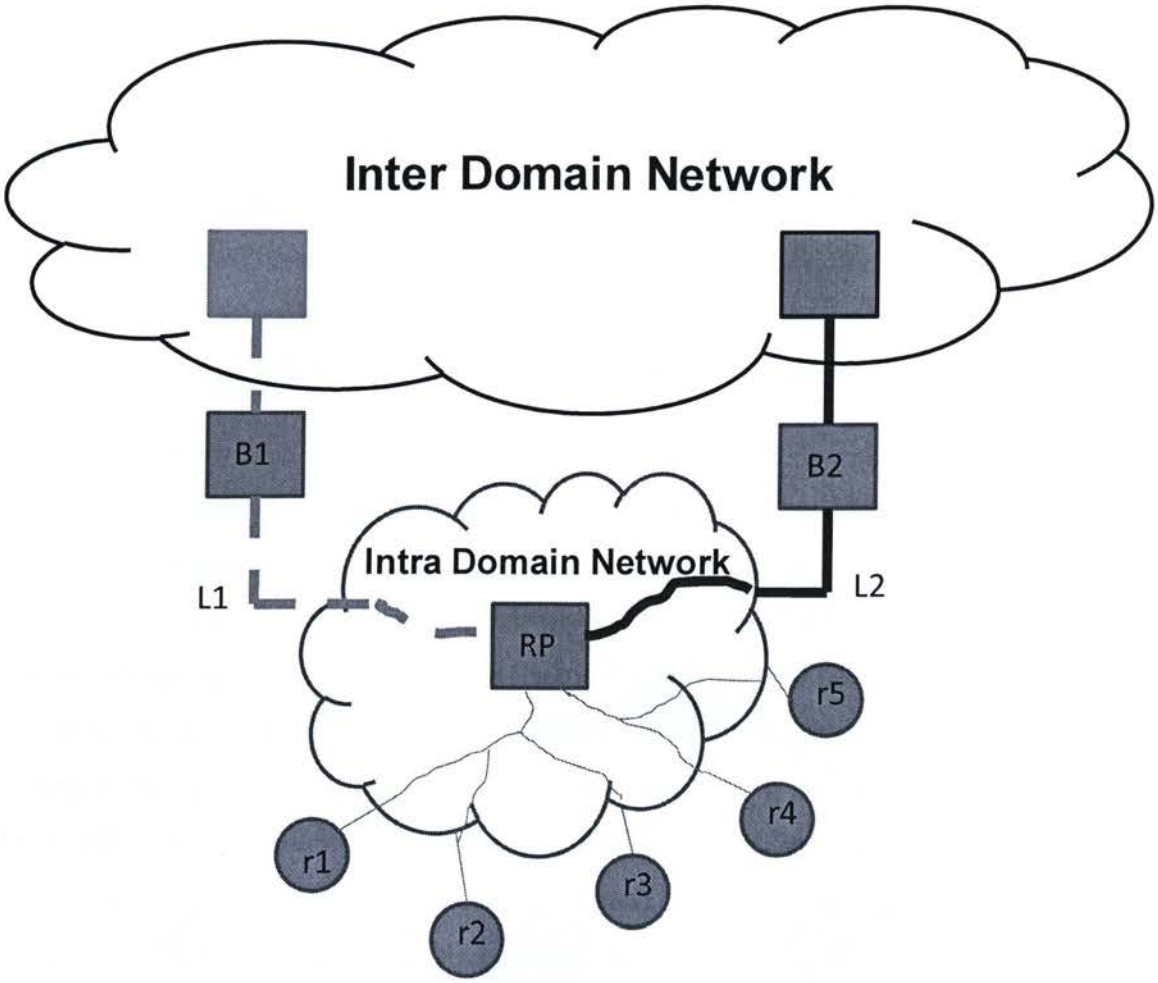

Figure 3.5: Inter and Intra Domain Topology

Figure 3.5 outlines the basic structure of the overall topology. In the SRP model any router can be selected to act as a RP for each multicast stream. In comparison, the BRP model assumes that the borders routers are the RPs in which the same multicast stream is duplicated on both B1 and B2. This allows BRP to optimize intra domain bandwidth more efficiently then SRP, at the cost of consuming more bandwidth on the inter domain links. Figure 3.5 shows that a RP has been selected within the intra domain network. This RP has two possible paths to B1 or B2 which have been defined as L1 and L2. Only one of these links will be utilized for a multicast stream. For example, L1 may be selected and L2 will be designated as a backup path in case of link failure at B1. L2 can be assigned a higher link weight so that when the unicast table populates, L1 will be selected. This

procedure is repeated for each multicast stream. Each stream will have its own unique path and RP. This allows for optimization for each multicast stream.

To reduce this propagation delay of PIM Join the RP can be positioned inside the network such that it is located at a minimal hop distance between both B1 and B2. Since different RPs can be designated for different multicast streams in our SRP model, the SRP can be calculated for each multicast stream. Different multicast groups can assign the RP for a particular stream based on available bandwidth. This is especially important to avoid the over provisioning of a node (RP) with multiple multicast streams. Once an RP has been selected, the link optimization algorithm similar to that in BRP, can be executed to determine the multicast tree. BRl^{intra} in equation 3.3 defines the minimum path from one border router to another border router. BRl^{intra} quantifies the link cost between two border routers. By applying a minimization function, the cost can be reduced. A minimum hop count cost can be obtained by modeling the topology as a Steiner tree. The problem associated with this is that Steiner tree is a NP problem. The same solution can be employed as in the BRP which involves the use of the genetic algorithm to create different solution sets. D_i represents the demand of bandwidth for a particular link in the topology. BRT_i represents the path between two different border routers. BRl^{intra} sums all links on the optimized path between two border routers. For example Figure 3.5 would have a BRT_i of 2 [3]. Once the path has been defined by the algorithm the link weights are altered so that when the routing table is populated it reflects the optimized path. To avoid issues affecting other network services, link weights can be applied to a particular multicast stream.

$$\text{minimize } BRl^{intra} = \sum_{i=1}^{t} \sum_{(u,v)} D_i \times x_{uv}^{i}$$

$$where\ x_{uv}^{i} = \begin{cases} 1\ if\ (u,v) \in BRT_i \\ 0\ otherwise \end{cases} \quad \text{.....Equation 3.3}$$

The SRP algorithm is defined below in algorithm 3.2. The algorithm is similar to algorithm 3.1 which is the basis of BRP. The SRP algorithm first defines a RP point. This point will determine the root node of the multicast stream. By picking a central point between two border routers this ensures that the hop count between each border router and the RP are equidistance.

Procedure SRP

Begin

Set the *M-IGP* weight of each intra-domain link in the network according to the chromosome;

For each prefix P_j

Aggregate group bandwidth demand according to P_j, i.e.,

$$AD_j^{\text{inter}} = \sum_{i=1}^{I} D_i \text{ for } s_i \in P_j;$$

End for;

Sort the prefix list P in descending order according to $AD_j^{\text{inter}} (0 < j < k)$;

While RP = {};

minimize $\text{BR}l^{\text{intra}}$

If $\text{BR}l^{\text{intra}} \left(\frac{|\text{BR}l^{\text{intra}}|}{2}\right)$ has available bandwidth

Assign an RP based upon the following condition $\text{BR}l^{\text{intra}} \left(\frac{|\text{BR}l^{\text{intra}}|}{2}\right)$, where *RP* has sufficient residual bandwidth for selection

End If;

End While;

For each prefix P_j in the ordered list P

intra-domain bandwidth consumption $l^{\text{intra}}\left(\{RP\}\right)$ is minimized for the groups whose source $s_i \in P_j$ and ***RP*** has sufficient residual bandwidth for the aggregated

demand AD_j^{inter};

Update inter-domain link utilization on ***RP***, i.e.,

$$u_b^{\text{inter}} = u_b^{\text{inter}} + \frac{AD_b^{\text{inter}}}{C_b};$$

End for;

$l^{\text{inter}} = \sum_{j=1}^{k} l_j^{\text{inter}};$ /* Sum up total intra-domain bandwidth consumption for all prefixes*/

$$fitness = \frac{\alpha}{l^{\text{intra}} + \alpha \times \max(u^{\text{inter}})};$$

End

Algorithm 3.2: SRP Algorithm

Once the RP is selected optimization can occur from the RP to the receivers. Each multicast group repeats the same procedure. Every iteration performs a bandwidth check which ensures that over provisioning does not occur. The remainder of the SRP algorithm is same as the BRP algorithm in algorithm 3.1.

3.5 Modified Dijkstra algorithm for link optimization in SRP and BRP

The Dijkstra algorithm was modified so that it produced a minimal hop count optimization rather than the shortest path optimization. This was used in both SRP and BRP to minimize the bandwidth consumption. The modified Dijkstra algorithm determines the shortest path between two points; for example the RP and a receiver. This modified Dijkstra algorithm is applied to all the receivers and to the RP. The receiver which returns the least link cost will be selected as the main path and all routers along that path will be marked. The next receiver will compare its link cost from all marked routers to its location in the topology.

An example of how this algorithm is applied to SRP can be seen in figure 3.6. The dashed lines represent physical links that are not selected. The solid lines represent activated links. B1 and B2 represent border routers on the intra-domain side. The routers are defined as R1 to R3 and the receivers have been defined as r. A link weight of 20 is then assigned to all non activated links. This forces the multicast stream to follow the defined optimized path.

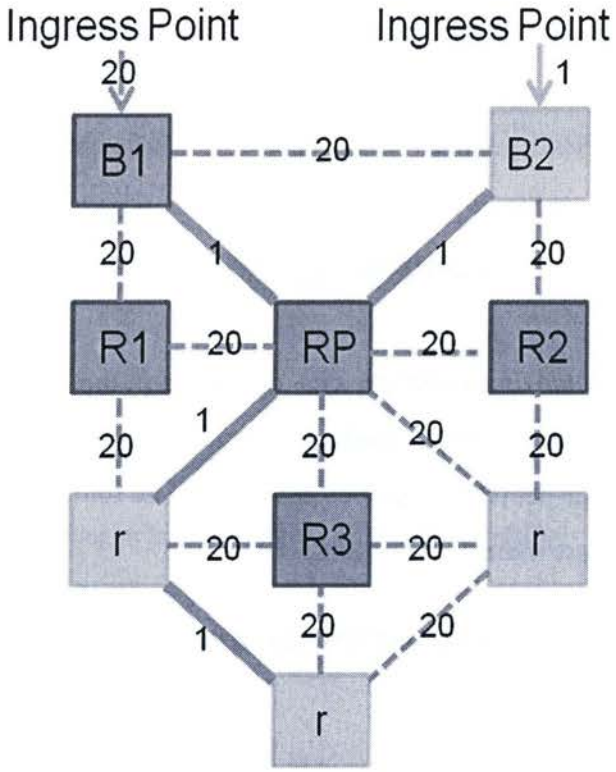

Figure 3.6: Modified Dijkstra's algorithm

Figure 3.7 represents how the original Dijkstra algorithm which would calculate the shortest path to all the receivers. From the two different figures two distinct paths can be seen. In figure 3.6 the total links utilized is 4 whereas in figure 3.7 there are 6 utilized links. This simple example demonstrates the key differences in the two algorithms.

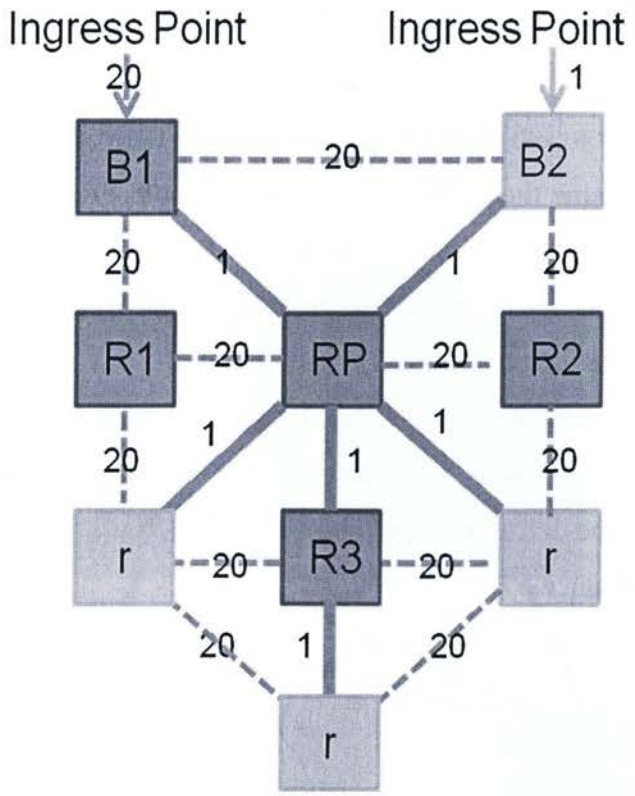

Figure 3.7: Applied Dijkstra's algorithm

The RP is selected by different factors such as link utilization, and shortest path. This algorithm represents how the results were optimized in the simulation by considering the shortest path. To ensure that a RP is selected, links that directly connected border routers are not considered in the optimization algorithm. This is to ensure that the border routers are not selected as an RP for the SRP algorithm.

Chapter 4: Simulation and Analysis

Simulation was used to evaluate the performance of SRP and BRP under link failure conditions. To determine their relative performance, both SRP and BRP were simulated under the same network conditions. The selected simulator was NS2 because of its robustness and support for the multicast protocol. To define the topology, link weights were used to force the optimized unicast routing table. These topologies were allowed to reach a steady state in NS2. Once the topology was defined, a scripted link failure was triggered and the results were then captured. In this chapter the parameters and the implementation will be discussed regarding BRP and SRP in relation to the simulator and their relative performance.

4.1 Simulation Set-up

Random topologies were created using BRITE with a Waxman distribution model. Each random topology consisted of node counts starting from 25 nodes and incrementing by 25 nodes until 100 nodes were achieved. Each node set was further subdivided into varying receiver densities which were incremented from 5% receiver density to 60% receiver density in increments of 5%. Receivers were represented as routers in the simulation; this mimicked the actual representation in the SRP and BRP algorithm. This procedure was repeated 20 times per receiver density with a random topology. Figure 4.1 outlines the structure of the topology generation. Only two border routers were used in the simulation. In order to create different multicast distribution trees, corresponding to different multicast groups in a single topology in NS2, link weights were used to force the optimized path. This method has been discussed in [10] using Multi Topology OSPF (MT-OSPF). This allows for different topologies to be formulated for different network services, which can be used to create multicast distribution trees for each multicast group. NS2 has not implemented this protocol; for this reason each topology was simulated separately with NS2. These topologies were then analyzed and condensed.

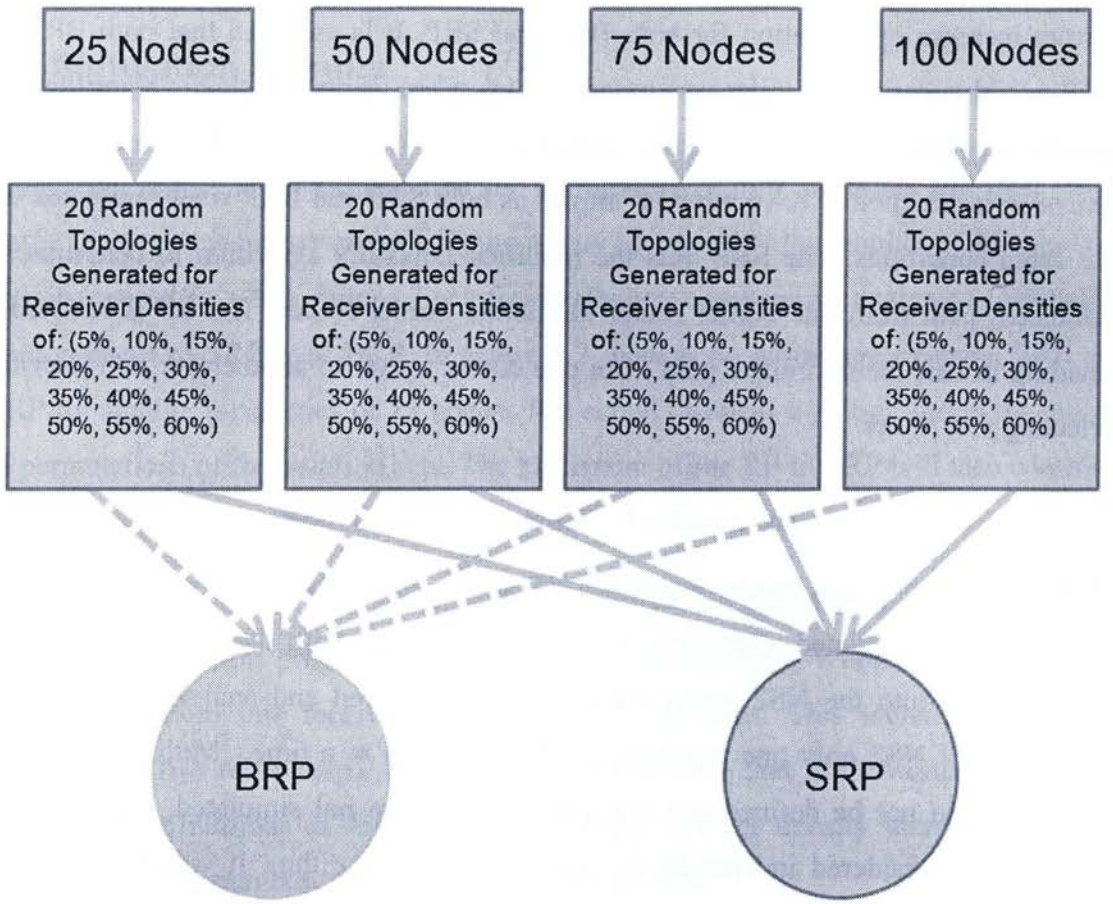

Figure 4.1: Topology Generation Structure

SRP and BRP were applied to each topology. These algorithms determined the optimal path; a link weight of 1 was applied for each link along the optimized path. A higher link weight of 30 was assigned to all links that were not considered optimized. The upper bound link weight was determined based on the topology size. Initially a random topology was generated with different receivers to node densities. For SRP and BRP the shortest path was found by exploring each node and by using a modified Dijkstra's algorithm. While not as efficient as genetic algorithm, it provided similar results. This is attributed to the fact that the genetic algorithm does not determine the optimal path; rather it creates new combinations of link weight sets which would then be processed by

the modified Dijkstra's algorithm. Both SRP and BRP are subjected to the same optimization of the modified Dijkstra's algorithm. Using the genetic algorithm will only further increase optimization for both BRP and SRP. It is assumed that both BRP and SRP would experience similar performance gains when using the genetic algorithm. The genetic algorithm was omitted from simulation to reduce the overall complexity. This omission is assumed to have minimal impact as both SRP and BRP were optimized with the same conditions. The SRP uses the modified Dijkstra's algorithm to determine the optimized path from one border router to another border router. The RP is selected by dividing in half the optimized path from one border router to another border router then selecting that router.

BRP assumes that each border router is a RP and performs the same optimization. Once this has been completed a tcl scrip file will be generated with the optimized topology. This is input into the NS2 where the results are captured and analyzed. Due to the limitations of NS2 only one topology can be simulated at a time. Multi topology link weights could not be defined and for that reason where not simulated. Each multicast stream was considered independently throughout the algorithm. It was assumed since both BRP and SRP have been subjected to the same conditions the results will be valid. Two different benchmarks were used to evaluate the relative performance of SRP and BRP; the first was the number of PIM Join messages. By monitoring the number of PIM Join messages it can be determined by how much the topology had shifted. The second benchmark was link traffic density surrounding the border routers. This indicates the amount of congestion formed by sending duplicate multicast streams on the links surrounding the border router. For example, if BRP optimized the border router to send the same multicast stream on 5 of its 5 links this would be then compared to SRP which could be optimized to only use 2 of the 5 links. PIM Join messages and interconnecting routers were tabulated based upon a per node basis. Time stamps from the NS2 output were used to remove the initial Join messages for both SRP and BRP. By removing the

initial Join messages, only Join messages that are related to the link failure can be analyzed.

4.2 Simulation Results

Simulation results are presented first for the number of join messages and then for the traffic density near border routers. Figure 4.2 to 4.5 represent the number of Join message activity caused by a link failure. Each figure shows a comparative analysis of percentage improvement of SRP over BRP. In figure 4.2 there is no advantage of using SRP over BRP for smaller topologies of 25 nodes. The actual physical topology limits the number of permutations of the solution sets. The placement of the RP in SRP will also contribute to higher hop count compared to a more optimized solution, such as BRP.

As the number of nodes increase from figure 4.2 to 4.5 it increases the possible solution set. The node count is proportional to the span of the topology. When the span increases, the distance from one receiver to another also increases. This increase in the span between any two receivers or, from a receiver to a node that is forwarding the stream, determines the number of Join messages. As the receiver density increases, the span decreases and the number of Join messages also decrease.

While Join and Prune messages cause small control traffic, they are responsible for initializing large streams of data. A Typical MP4 stream of HD video for one channel can require up to 19 Mbps [17]. When multiple channels are considered the bandwidth requirements are even higher. Typical IPTV providers can provide hundreds of channels.

Reduced Join messages within a topology are indicative of a reduced shift in traffic. Both BRP and SRP consider conservation of available bandwidth. With BRP, during a link failure, the links are not defaulted to an optimized solution where available bandwidth could be considered. The SRP includes two optimized paths: a standard operational path and a failure path. This, in conjunction with the placement of the RP, reduces the system stress during a link failure. Join message concentration occurs from the RP to the receiving routers when using SRP. This minimizes the number of Join messages and the

shift in topology when a link failure occurs. When BRP is used, Join messages occur from all potential links that are connected with the original border router.

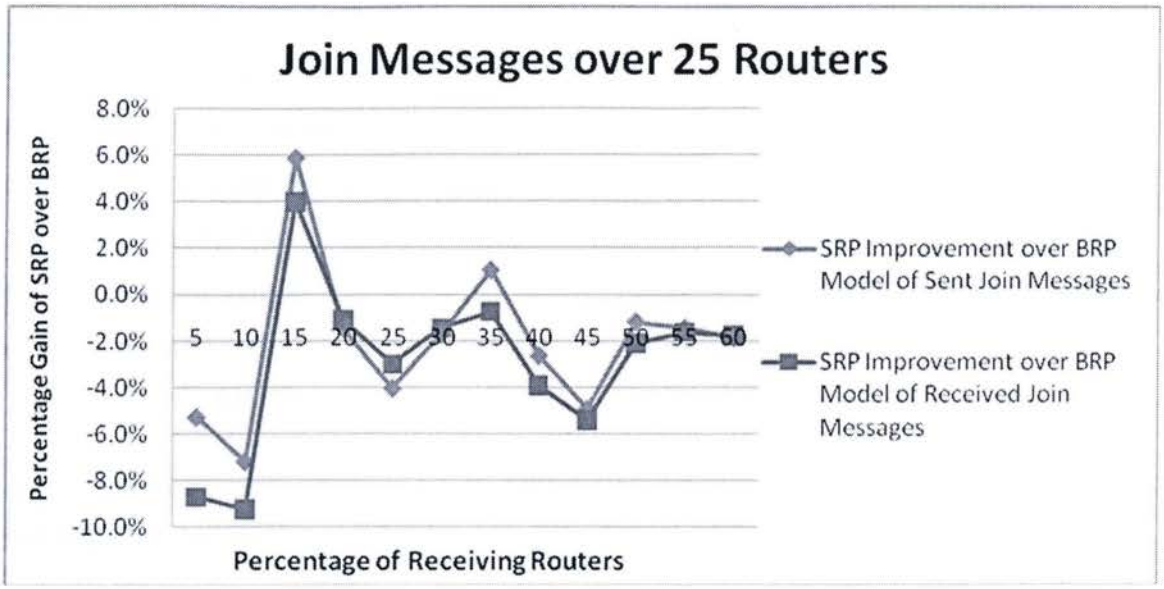

Figure 4.2: SRP Improvement over 25 Nodes

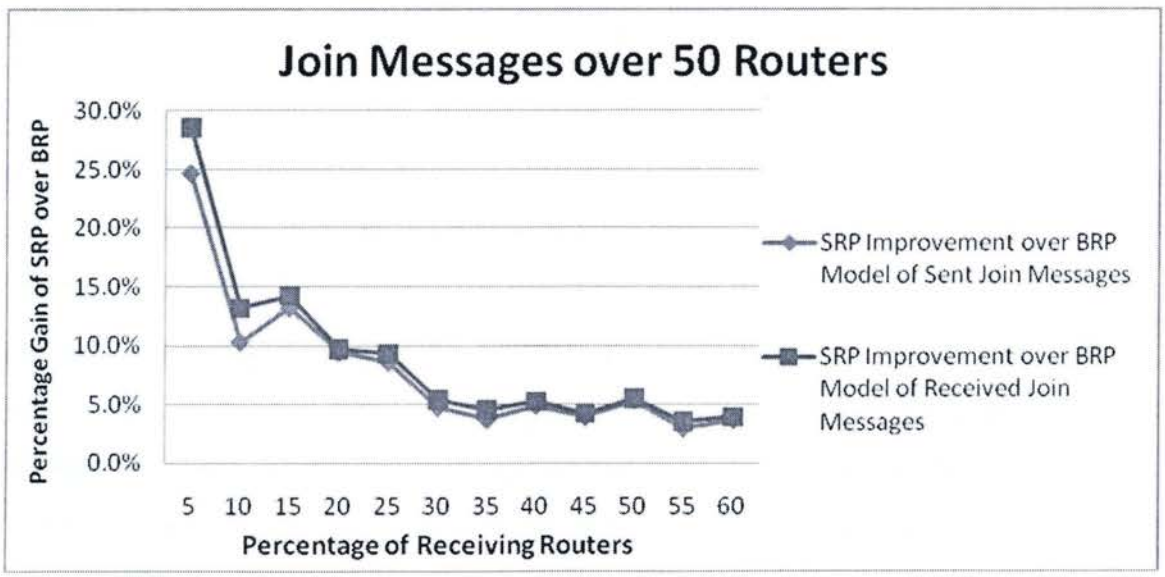

Figure 4.3: SRP Improvement over 50 Nodes

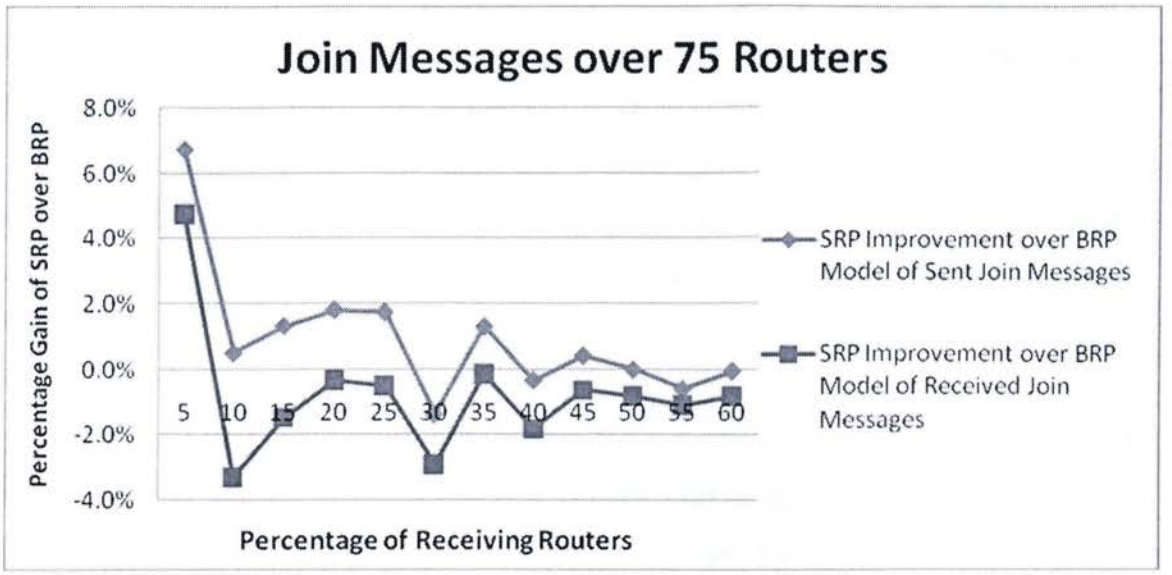

Figure 4.4: SRP Improvement over 75 Nodes

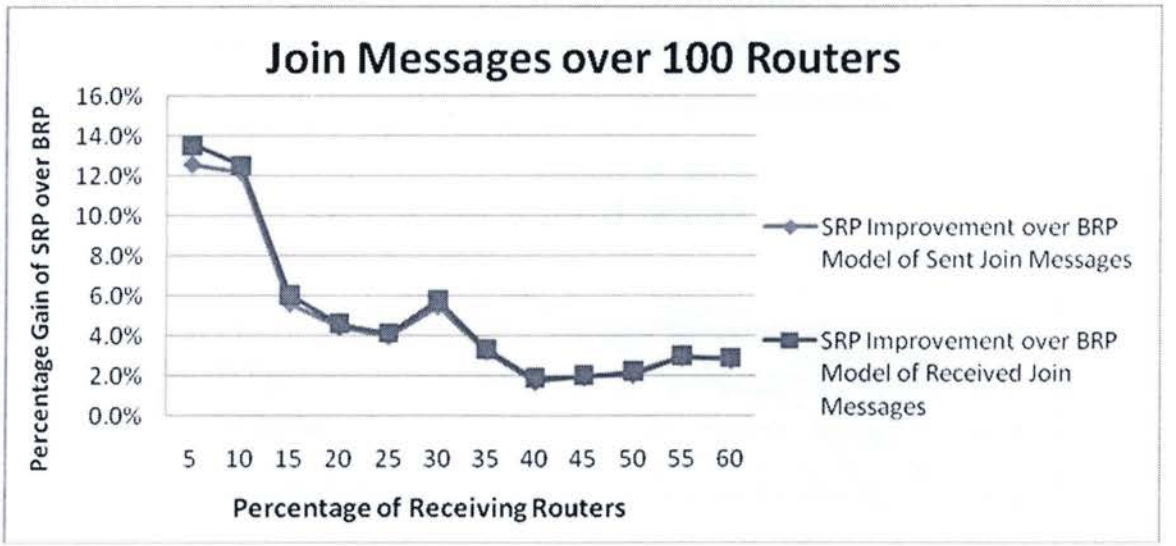

Figure 4.5: SRP Improvement over 100 Nodes

The measure of active links involved in carrying the multicast stream surrounding a border router indicates the concentration of traffic at, and around, the border router. Figure 4.6 compares percentage improvement of these actives links for SRP with BRP.

As the percentage of receivers increases in a topology the effectiveness of SRP becomes more pronounced. SRP compared to the BRP algorithm shifts the root of the multicast stream from the border router to a centralized RP within the intra-domain environment. This removes link congestion surrounding the border router and allows other traffic to utilize the spared bandwidth. Congestion on the links closest to the border router will be reduced. BRP does not take into consideration link congestion surrounding the border router and for this reason has a higher active link density surrounding the border routers.

Figure 4.6 for router set of 25 shows negative results when using SRP for low receiver densities. This is attributed to the locality of the receivers. As the receiver density increases the locality of the receivers in relation to the border router determines the number of links required. BRP optimization without a central RP causes congestion surrounding the border routers when considering high receiver densities. Figure 4.6 illustrates the effectiveness of SRP improved as the node count increases. This is directly attributed to the increase of nodes and their relative position in relation to the border routers.

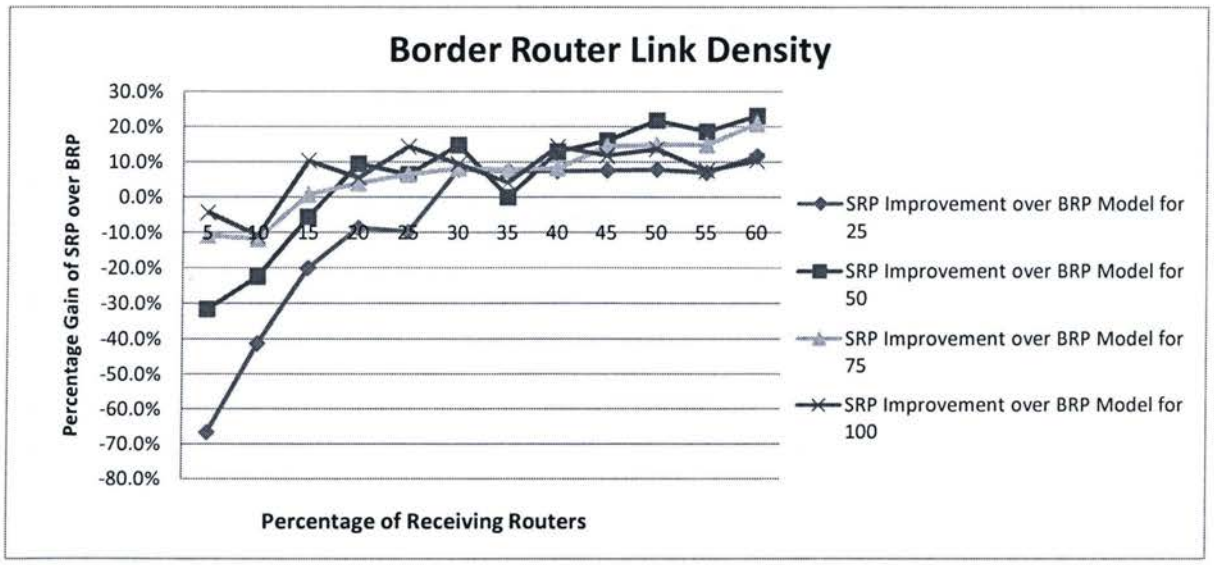

Figure 4.6: Active Link Density at Border Router

The results of SRP for active link density and Join messages traffic show that they are inversely proportional to each other. As active link density performance increases, Join message performance decreases. From the results reviewed, 30 percent receiver density provides the most benefit when considered jointly with the improvement in link density and Join message traffic. Shifts in topology which has been quantified by Figure 4.2 to 4.5 affect intra domain traffic whereas border router link density affects inter domain traffic.

SRP displays gains compared to BRP in reducing traffic and congestion during a link failure within the intra domain environment. SRP also has another advantage over BRP. With BRP, it optimizes hop counts to reduce link congestion by means of using multiple border routers in the simulation 2 border routers were used. The total available inter domain bandwidth when using BRP is reduced for the topology. This is primarily due to redundancy that is created by the BRP algorithm. SRP however calculates two paths towards the border router but only initializes one of them. This removes the redundancy that is created with the BRP algorithm. By removing this redundancy it also reduces the inter domain bandwidth demand.

Figures 4.2 to figure 4.5 indicate varying amounts of improvement using the SRP algorithm over BRP algorithm. The inconsistency in the results could be attributed to the limited data set or random sets of receivers. One possible solution is to increase the number of randomly generated topologies for each receiver density. Currently there are only 20 topologies per receiver density. By increasing the number of topologies that are simulated more combinations will be evaluated which should reduce the inconsistencies within the figures. Even with the irregularities there is still a positive trend which is clearly visible across all graphs for both Join messages activity and border router link density.

Chapter 5: Conclusion

There is a resurgence of interest in multicast due to its bandwidth efficiency in distributing IPTV content. While the demand for IPTV content is increasing, so does the need for reliability and efficiency. While the efficiency of such deployment has been studied, there is a lack of understanding of how the reliability and efficiency are correlated. Reliability is an important factor in any network topology. SRP tries to correlate both reliability of exploiting multiple ingress border routers and efficiency of reducing the number of actual links required to service a receiver set. Both SRP and BRP take advantage of multiple border routers. As bandwidth costs continue to erode egress links to the inter domain become more affordable.

IPTV with HD content can consume up to 19 Mbps per stream. An IPTV provider network typically carries hundreds of these streams. These providers also offer other services to consumers within their network. These services include high speed internet and VOIP. The SRP can be deployed to reduce congestion surrounding border routers and also congestion within the topology during a link failure. A sudden change in a topology could cause over provisioning of links within the intra domain topology. Typically high speed internet is not a high priority service for an ISP compared to VOIP. Hence, web traffic is usually not prioritized in the network. Without SRP a sudden change in topology could affect other services including both multicast traffic and web traffic. The SRP also limits one ingress border router per multicast stream compared to BRP, which can assign multiple border routers per multicast stream. By assigning multiple border routers per stream any bandwidth advantage will be reduced. In contrast SRP defines the path towards the second border router but does not does not use it until the first border router becomes inaccessible. This ensures if a link failure does occur on the primary border router than the second path can be initialized quickly. This second path has been optimized to reduce the hop count from the RP towards the second border router. By configuring the multicast stream in this manner inter domain bandwidth can

be conserved. The SRP also removes link congestion surrounding the border router by reducing the number of links per multicast stream. This is achieved by changing the distribution of the multicast stream to the RP.

Bandwidth conservation continues to be an important topic as consumers are exploring more multimedia applications. Network providers traditionally avoid costly upgrades and instead rely on other measures such as Deep Level Packet Inspection to control bandwidth within their networks. The SRP could be used as a tool to assist in avoiding costly upgrades and increase the efficiency of a network.

Future work concerning SRP can incorporate multiple ingress points and balancing link failures between multiple sets of ingress points. This would allow for even better control of bandwidth during a link failure. The effects of route oscillations during heavy traffic loading are another area to explore. The SRP provides an improvement over BRP in network congestion control. All these topics will allow for greater insight of the effectiveness of SRP and they should be explored in the near future.

References

[1] Wielosz, Anna.; Islam, Kashif. "Achieving fast restoration times in IP networks for IPTV video transport" – case study. Cisco Systems, Inc.

[2] Wang, N.; Pavlou, G. "Traffic engineered multicast content delivery without MPLS overlay Multimedia", IEEE Transactions on Volume 9, Issue 3, April 2007 Page(s):619 - 628

[3] Wang, N.; Pavlou, G. "An efficient IP based approach for multicast routing optimisation in multi-homing environments. Next Generation Internet Design and Engineering", 2006. NGI apos;06. 2006 2nd Conference on Volume , Issue , 3-5 April 2006 Page(s): 8 pp.

[4] T.C. Bressoud et al. "Optimal Configuration for BGP Route Selection", IEEE INFOCOM, 2003.

[5] T. Przygienda et al. " M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems". Cisco Systems, February, 2008

[6] P. Psenak et al, "Multi-Topology (MT) Routing in OSPF" Internet Draft, draft-ietf-ospf-mt-04.txt Apr. 2005

[7] B. Fortz et al, "Internet Traffic Engineering by Optimising OSPF Weights", IEEE INFOCOM, 2000, pp. 519-528

[8] A. Sridharan et al, "Achieving Near-Optimal Traffic Engineering Solutions for Current OSPF/IS-IS Networks", IEEE INFOCOM, pp. 1167-1177, Apr. 2003

[9] Y. Wang et al, "Internet Traffic Engineering without Full Mesh Overlaying", Proc. IEEE INFOCOM, Vol. 1, pp. 565-571, 2001

[10] N. Wang, G. Pavlou, "Bandwidth Constrained IP Multicast Traffic Engineering without MPLS Overlay", IEEE/IFIP MMNS, 2004

[11] The GEANT network topology, available online at: http://www.geant.net/upload/pdf/Topology_Oct_2004.pdf

[12] Wong,T.;Van Jacobson;Alaettinoglu, C. "Internet routing anomaly detection and visualization". Dependable Systems and Networks, 2005. DSN 2005. Proceedings. International Conference on 28 June-1 July 2005 Page(s):172 – 181

[13] Zhang, Jian; Rexford, Jennifer; Feigenbaum, Joan. "Learning-Based Anomaly Detection in BGP Updates". Proceeding of the 2005 ACM SIGCOMM workshop on Mining network data, 2005.

[14] P. Rajvaidya, K. Almeroth."Multicast Routing Instabilities", IEEE Internet Computing, Vol. 8, Issue 5, 2004, pp. 42-49

[15] Haakon Ringberg; Augustin Soule; Jennifer Rexford; Christophe Diot. "Sensitivity of PCA for traffic anomaly detection". SIGMETRICS '07 Conference Proceedings, June 2007

[16] Soule, Augustin; Salmatian, Kave; Taft, Nina. "Combining Filtering and Statistical Methods For Anomaly Detection". IMC '05, 2005 Internet Measurement Conference , Pp. 331–344

[17] Simsarian, J.E.; Duelk, M. "IPTV Bandwidth Demands in Metropolitan Area Networks". Local & Metropolitan Area Networks, 2007. LANMAN 2007. 15th IEEE Workshop on 10-13 June 2007 Page(s):31 – 36

[18] Ying-Dar Lin; Nai-Bin Hsu; Chen-Ju Pan.;"Extension of RP relocation to PIM-SM multicast routing". Communications, 2001. ICC 2001. IEEE International Conference on Volume 1, 11-14 June 2001 Page(s):234 - 238 vol.1